Multivariate Data Analysis with Applications to Cancer

Dissertation Advisor: Professor Yi Li Anna Catherine Snavely

Multivariate Data Analysis with Applications to Cancer

Abstract

Multivariate data is common in a wide range of settings. As data structures become increasingly complex, additional statistical tools are required to perform proper analyses. In this dissertation we develop and evaluate methods for the analysis of multivariate data generated from cancer trials. In the first chapter we consider the analysis of clustered survival data that can arise from multicenter clinical trials. In particular, we review and compare marginal and conditional models numerically through simulations and discuss model selection techniques. A multicenter clinical trial of children with acute lymphoblastic leukemia is used to illustrate the findings. The second and third chapters both address the setting where multiple outcomes are collected when the outcome of interest cannot be measured directly. A head and neck cancer trial in which multiple outcomes were collected to measure dysphagia was the particular motivation for this part of the dissertation. Specifically, in the second chapter we propose a semiparametric latent variable transformation model that incorporates measurable outcomes of mixed types, including censored outcomes. This method extends traditional approaches by allowing the relationship between the measurable outcomes and latent variable to be unspecified, rendering more robust inference. Using this approach we can directly estimate the treatment (or other covariate) effect on the unobserved latent variable, enhancing interpretation. In the third chapter, the basic model from the second chapter is maintained, but additional parametric assumptions are made. This model still has the advantages of allowing for censored measurable outcomes

and being able to estimate a treatment effect on the latent variable, but has the added advantage of good performance in a small data set. Together the methods proposed in the second and third chapters provide a comprehensive approach for the analysis of complex multiple outcomes data.

Contents

List of Figures

List of Tables

Acknowledgments

I want to take this opportunity to acknowledge a number of people who have contributed to this dissertation. First, I would like to thank Yi Li for working with me over the past several years and for providing me with countless opportunities to grow as a statistician. In addition, I want to thank Xihong Lin and David Christiani for serving on my committee and providing useful insight and suggestions. Finally, I would like to express my great appreciation to Dave Harrington for all of his advice and support. Dave, I will always be grateful to you for making time for me and for believing in me.

This dissertation would not have been possible without having data to analyze and interesting clinical questions to answer. Therefore, I would like thank Jim Anderson for providing the COG data used in the first chapter and Drs. Laura Goguen and Claudia Chapuy for providing the motivation and data for the second and third chapters.

I have enjoyed my time at Harvard and that is a direct result of the faculty, staff, and students in the Biostatistics Department. I would like to thank the faculty for providing me with a solid statistical foundation and the staff for keeping the department going. A special thanks to Jelena Follweiler for always looking out for me. To my classmates and friends, thanks for making my time in Boston a little more fun.

Most importantly, I would like to thank my family for their unconditional love and support. I am so lucky to have such an amazing group of people in my life.

A Comparison of Methods for Analyzing Clustered Survival Data in Multicenter Clinical Trials

Anna C. Snavely and Yi Li

1.1 Abstract

Clustered survival data often arise from large randomized clinical trials conducted in multiple centers. There are two major classes of models for addressing clustering in this setting: marginal (population-averaged) models and conditional (center-specific) models. This paper reviews and compares marginal and conditional models (frailty models in particular) numerically through simulations that consider the impact of model misspecification on point estimates for both true marginal models and true frailty models. We show that large differences can exist between marginal and conditional coefficients, particularly when the within cluster dependence is strong. We also demonstrate that AIC/BIC cannot be used for model selection.

1.2 Introduction

Many large randomized clinical trials are carried out at multiple medical centers in order to facilitate the recruitment of a sufficient number of patients. Multicenter designs also allow for greater generalizability of results. However, additional complications arise with multicenter designs. Different medical centers tend to have different patient populations, different doctors, and different standard practices. Because of these center differences, it is likely that patients within the same center are more similar to each other than to patients in a different center even though trial-specific protocols are followed. This can lead to dependence, or clustering, of outcomes from the same center. See, for example, Anello et al. (2005), Fleiss (1986), Gray (1994), and Senn (1998). Therefore, when doing an analysis of such a trial, it is important to take this clustering into account. When the effect of clustering is strong, not accounting for the dependence can lead to misleading inference (Glidden & Vittinghoff, 2004).

A common outcome in multicenter clinical trials is survival or some other failure time. This makes exploring methods for accounting for clustering in time-to-event data important. This paper reviews some of the available methods for analyzing such clustered time-to-event data and presents simulation results comparing the various methods. In particular, the interest of many randomized trials is assessing a potential treatment effect. However, this treatment effect can have either a population-averaged interpretation or a conditional interpretation depending on which method of analysis is used. Therefore, this paper seeks not only to compare the different methods, but also to explore relationships between population-averaged and conditional effects and to consider how to choose between the various models. Simulations investigate varying levels of censoring and correlation within center, and varying number of clusters. The methods are also compared using data from a Children's Oncology Group multicenter trial for acute lymphoblastic leukemia.

1.3 Methods For Analyzing Clustered Failure Time Data

Suppose censored failure time data is obtained from a multicenter clinical trial with J clusters (centers) and with n_j subjects in cluster j $(j = 1, \ldots, J)$. The total sample size is then $N = \sum_j n_j$. Let T_{ij} and C_{ij} be the failure time and censoring time for subject i in cluster j. We then observe $X_{ij} = \min(T_{ij}, C_{ij})$, the follow-up time, and $\Delta_{ij} = I(T_{ij} < C_{ij})$, the failure indicator. Z_{ij} is a vector of covariates for subject i in cluster j. Let $\mathbf{T}_j$ be the vector of failure times for cluster j, $\mathbf{C}_j$ be the vector of censoring times for cluster j and $\mathbf{Z}_j$ be the covariate matrix for cluster j. Then, we assume that $\mathbf{T}_j$, $\mathbf{C}_j$, and $\mathbf{Z}_j$ are independent across centers and that $(\mathbf{T}_j, \mathbf{C}_j)$ are conditionally independent given $\mathbf{Z}_j$.

When dealing with clustered time-to-event data, there are two major classes

of models: marginal (or population-averaged) models (Wei et al., 1989; Lin, 1994; Prentice & Cai, 1992; Cai & Prentice, 1995) and conditional (center-specific) models (Murphy, 1994, 1995; Parner, 1998; Vaupel et al., 1979). If we consider a Cox model with covariate vector Z_{ij}, the marginal hazard for subject i in center j is:

$$\lambda_{ij}(t|Z_{ij}) = \lambda_0(t)\exp(\beta^T Z_{ij}). \tag{1.1}$$

In this model, the baseline hazard, $\lambda_0(t)$, is not specific to a particular center. If we consider the interpretation of the coefficient (β_T) for a treatment covariate in this setting (1 indicating treatment and 0 indicating placebo), e^{β_T} is the marginal hazard ratio that compares the risk of failure for an individual who receives treatment and an individual who receives placebo who are randomly selected from the population. This means that marginal models give a population-averaged interpretation for the regression parameters. Both the naive (traditional) Cox model and the marginal Cox model (takes into account clustering) produce coefficients with population-averaged interpretations.

On the other hand, if the center is explicitly taken into account when modeling the hazard, we get a conditional model of the form:

$$\lambda_{ij}(t|Z_{ij}) = \lambda_{0j}(t)\exp(\beta^T Z_{ij}). \tag{1.2}$$

Center is included in this model through a center-specific baseline hazard, $\lambda_{0j}(t)$. In this case, e^{β_T} is the hazard ratio that compares the risk of failure for an individual who receives treatment and an individual who receives placebo who are from the same center. In other words, regression parameters in conditional models must be interpreted as conditional on the center (Glidden & Vittinghoff, 2004). The fixed effects Cox model, which is fit by including dummy variables for each center in the traditional Cox model (Glidden & Vittinghoff, 2004), the stratified Cox model (Holt & Prentice, 1974), and the frailty model are all examples of conditional models. Frailty models are particularly useful and will be the only class of conditional models considered further in this paper.

Since the βs in model (1.1) and model (1.2) have different interpretations, in general it is not expected that the values for the two coefficients would be the same. In fact, the only time we would expect the coefficients to be the same in the marginal and conditional models is when there is no correlation within the center (Therneau & Grambsch, 2000).

1.3.1 The Naive Cox Model (A Population-Averaged Model)

The Cox proportional hazards model (Cox, 1972, 1975) is the most popular method for analyzing failure time data. Model (1.1) represents the form of the Cox model, where $\lambda_0(t)$ is an unspecified non-negative function. Estimation of β is based on the partial likelihood:

$$L(\beta) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} \left[\frac{\exp(\beta^T Z_{ij})}{\sum_{k=1}^{J} \sum_{l=1}^{n_k} I(X_{kl} \geq X_{ij}) \exp(\beta^T Z_{kl})} \right]^{\Delta_{ij}}. \tag{1.3}$$

The partial likelihood is not a full likelihood, but can be treated as such for the purpose of inference. This means the estimate of β can be found by solving the score equation and the inverse information can be used to estimate the variance. The Cox model assumes that the observations are independent. However, this assumption does not hold in the case of clustered failure time data. Therefore, if the naive Cox model is used for clustered time-to-event data, variance estimates are likely to be too small (Lorino et al., 2004). As such, when your data are truly clustered the naive Cox model should not be used for analysis.

1.3.2 The Marginal Cox Model (A Population-Averaged Model)

The estimates for the regression parameters in the naive Cox model are fine when we have clustered failure time data. This means that parameter estimates can still be obtained using the partial likelihood. However, the variance term needs to be

corrected in an analogous manner as is done in the generalized estimating equation approach of Liang et al. (1992).

One approach to getting a robust variance estimator is by using the grouped jackknife. This means that the model is fit leaving out one cluster at a time, giving us jackknife influence values: $\hat{\beta}_{(j)} - \hat{\beta}$. These jackknife influence values can then be used to estimate the variance (Therneau & Grambsch, 2000).

Another approach is to use an approximation to the jackknife in the form of a sandwich estimator: ABA. In this estimator, A is the usual variance estimate and B is a correction term. In our setting the sandwich estimator can be written as $V = I^{-1}(U^T U)I^{-1}$, where I is the observed information matrix and U is the matrix of score residuals (Therneau & Grambsch, 2000). Lin & Wei (1989) developed an appropriate sandwich estimator for the Cox model which is algebraically equivalent to V. They showed this estimate is consistent and robust to several forms of misspecification.

A final approach is to use a modified sandwich estimator (Therneau & Grambsch, 2000; Lee et al., 1992). This estimator can be written as $V^* = I^{-1}(\tilde{U}^T\tilde{U})I^{-1}$, where $\tilde{U}$ is the collapsed score matrix obtained by replacing each cluster of rows in U by the sum of those rows. This is the approach that R and S-Plus use to get a robust variance estimator (Lorino et al., 2004).

1.3.3 The Frailty Model (A Conditional Model)

The frailty (or random effects) model assumes that center has a proportional effect on the baseline hazard function, that the center effects come from some probability distribution and that there is a constant treatment effect across centers. The random center effects are continuous variables that describe the excess risk (frailty) for each center. The frailties account for center heterogeneity caused by unmeasured

variables (Aalen, 1988). In this formulation we assume that subjects in the same center have the same excess risk. Therefore, this formulation of a random effects survival model is often called a shared frailty model. The form of such a frailty model is:

$$\lambda_{ij}(t) = \lambda_0(t) w_j \exp(\beta^T Z_{ij}) = \lambda_0(t) \exp(\beta^T Z_{ij} + \gamma_j). \tag{1.4}$$

From this formulation we see that model (1.4) is just a special case of model (1.2) where the frailties (w_j) are assumed to act multiplicatively on a common baseline hazard, $\lambda_0(t)$ (Therneau & Grambsch, 2000). Two of the most common distributions for the w_js are gamma and log normal. The positive stable distribution is also sometimes used because of nice theoretical properties. An advantage of the frailty model is that the frailties (center effects) can be estimated. This is useful if differences between centers are of particular interest.

In the gamma frailty model (Murphy, 1994, 1995; Parner, 1998; Klein, 1992; Nielsen et al., 1992), w_j follows a Gamma$(1/\theta, 1/\theta)$ distribution. In the log normal frailty model (McGilchrist & Aisbett, 1991; McGilchrist, 1993) the γ_js follow a normal distribution with mean 0 and variance θ. Since the γ_js follow a normal distribution, the frailties (w_j) follow a log normal distribution. For both models, a larger θ means the frailties are more dispersed and there is a larger dependence within centers. Therefore, a larger θ leads to greater heterogeneity in the center-specific baseline hazards. If θ is 0, then the frailties are equal to 1, and failures are independent both within and across centers.

The penalized Cox model is an efficient approach of estimation for the gamma and log normal frailty models. This method of estimation proceeds by maximizing the penalized partial log-likelihood:

$$PPL = l(\beta, \gamma) - g(\gamma; \theta). \tag{1.5}$$

The first piece of (1.5) is the usual Cox partial likelihood and the second piece is a constraint function that penalizes less desirable values of w. The penalty function

for the gamma frailty model is $(1/\theta)\sum[\gamma_j - \exp(\gamma_j)]$ and the penalty function for the log normal frailty model is $(1/2\theta)\sum\gamma_j^2$.

For both the log normal and gamma frailty models we could consider averaging over the frailty distribution in order to see the effect of covariates on the marginal hazard. When this is done, it can be seen that both the gamma and the log normal models lead to non-proportional marginal hazards (Glidden & Vittinghoff, 2004; Lorino et al., 2004). On the other hand, when averaging over the frailty distribution in the positive stable frailty model (Hougaard, 1986; Fine et al., 2003), the proportionality of the marginal hazard is maintained. Because the proportionality is maintained, there is a direct relationship between the conditional and marginal coefficients. Let β be the coefficient in the positive stable frailty model and let γ be the coefficient in the marginal model. In this case the following relationship holds, where $0 \leq \alpha \leq 1$:

$$\gamma = \alpha\beta. \tag{1.6}$$

We can see from (1.6) that the conditional parameter from the frailty model is larger in magnitude than the marginal coefficient. Unlike the gamma and log normal frailty models, the positive stable frailty model cannot be fit using the penalized Cox model approach. The positive stable model can be fit using the EM algorithm, but this algorithm is quite slow and proper variance estimates require further computation (Therneau & Grambsch, 2000). A SAS macro by J. P. Klein implements the EM algorithm for the positive stable model, but no such software is available in S-Plus/R. Hougaard (2000) also has developed software for fitting the positive stable model, but it has not been made available to the public. While the positive stable model has nice theoretical properties, the lack of efficient publicly available software makes it less attractive. As such, the positive stable frailty model will not be considered in the simulation study.

1.4 Simulation Study

A simulation study was carried out with the aim of comparing and choosing between the naive Cox model, the marginal Cox model, the gamma frailty model and the log normal frailty model. The four models were compared both using data simulated to follow a proportional hazards model with a known marginal treatment effect and data simulated to follow a proportional hazards model with a known conditional treatment effect.

1.4.1 True Model: Marginal Model

In multicenter clinical trials, it is often the case that we are interested in the population-averaged interpretation of covariates (treatment in particular). This would suggest the use of the marginal Cox model. In order to evaluate and compare the performance of the naive Cox model and the marginal Cox model, data were generated to marginally follow a proportional hazards model with a known treatment effect. Under this simulation setting we consider estimates for the treatment effect (β), the model-based standard error, the empirical standard error, the bias, the mean squared error, the empirical coverage of 95 percent confidence intervals, AIC and BIC. We can fit the frailty models to the data as well and calculate the same quantities. However, because the marginal and conditional treatment effects are not expected to be the same, the bias (and by extension the MSE) represents the difference from the true marginal effect, not the difference from the true conditional effect. These simulations, therefore, do not let us see how well models estimate the true conditional effect. However, the study is useful to consider relationships between marginal and conditional coefficients across different settings, as well as to consider how the conditional coefficients compare to each other across different settings.

Method of Simulation

The clustered failure time data with a known marginal treatment effect were simulated using a normal transformation model approach (Othus, 2009; Li & Lin, 2006). If we assume a Cox proportional hazards model (1.1), then the survival function can be written as:

$$S(t) = \exp(-e^{\beta^T Z_{ij}} \Lambda_0(t)). \tag{1.7}$$

$\Lambda_0(t)$ in (1.7) is the cumulative hazard function. For simplicity we can take $\lambda_0(t) = 1$ and have only a single treatment (Bernoulli) covariate. A transformation and the probability integral transform can then be used to get T^*_{ij} random variables that follow a standard normal distribution:

$$T^*_{ij} = \Phi^{-1}(S(T_{ij})) = \Phi^{-1}(\exp(-e^{\beta^T Z_{ij}} T_{ij})) = \sqrt{\sigma} b_i + \epsilon_{ij} \tag{1.8}$$

Therefore, in order to generate clustered survival times (T_{ij}), we can generate T^*_{ij} by generating b_i from a N(0,1) distribution and ϵ_{ij} from a N(0,$1 - \sigma$) distribution and then transforming back.

Twenty-four different simulation settings were considered: 2 levels of censoring (30%, 60%), 4 levels of correlation (σ= 0, 0.25, 0.5, 0.75), and 3 levels of cluster number (30, 60, 90). For each setting, 500 simulations were carried out. The single treatment covariate was generated from a Bernoulli(0.5) distribution for each setting to give approximately equal numbers in the treatment and placebo groups, which mimics many clinical trials. The marginal treatment effect (β) was set to log(0.5) = -0.693 for each setting as well. Since we are primarily interested in the multicenter clinical trial setting, cluster size was randomly generated to allow clusters of varying size. Cluster size was constrained to be between 10 and 50 with a mean cluster size of 30, again to mimic the clinical trial setting.

Results

Table 1.1 presents simulation results for the models with 30% censoring where the true model is the marginal model and Table 1.2 presents similar results for models with 60% censoring. We will first focus on the two models with marginal interpretations. The first thing to note is that the estimates for β are the same under both the naive Cox model and the marginal Cox model, which implies the bias is also the same for both models. The fact that the estimates are the same is expected because the parameters in both cases are estimated using the partial likelihood (1.3). Because the same partial likelihood is used for both models, the AIC and BIC are also identical for the two models. The difference in the models, therefore, is in the standard errors. This difference in standard errors becomes greater as the correlation increases since the standard errors under the marginal model increase as the correlation increases. The naive model performs fairly well when the correlation is small ($\sigma = 0.25$). However, it is clear that the naive model has standard errors that are too small when the correlation within center becomes stronger. This can be seen by the somewhat poor empirical coverage of 95% confidence intervals for the higher levels of σ. For larger correlations, it also appears that while the marginal model performs much better than the naive model, the marginal model still tends to somewhat underestimate the standard errors as evidenced by the empirical standard errors being larger than the model-based standard errors.

In all the settings the marginal Cox model performs well. This can be seen by the small biases and the empirical coverage of the confidence intervals being close to 95% in all cases. However, the bias does tend to increase as the correlation increases. This pattern is more pronounced when the number of clusters is smaller. Also, when there is more correlation, the bias is smaller when there are more clusters. When the censoring is higher, the standard errors for both the naive Cox model and the marginal model are inflated compared to the corresponding

Table 1.1: Simulation results for models with 30% censoring (True model: marginal model)

	Clusters=30				Clusters=60				Clusters=90			
	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.50$	$\sigma = 0.75$	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.50$	$\sigma = 0.75$	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.50$	$\sigma = 0.75$
Naive Cox												
Mean β	-0.692	-0.696	-0.713	-0.723	-0.696	-0.702	-0.700	-0.708	-0.695	-0.691	-0.694	-0.702
Mean SE	0.102	0.102	0.102	0.102	0.072	0.072	0.072	0.072	0.058	0.058	0.059	0.058
Empirical SE	0.104	0.102	0.118	0.135	0.073	0.078	0.082	0.091	0.058	0.063	0.062	0.079
Bias	0.001	-0.003	-0.020	-0.030	-0.003	-0.009	-0.007	-0.015	-0.002	0.002	-0.001	-0.009
MSE	0.010	0.010	0.011	0.011	0.005	0.005	0.005	0.005	0.003	0.003	0.003	0.003
95% CI coverage	0.952	0.936	0.906	0.862	0.938	0.934	0.912	0.880	0.940	0.930	0.940	0.866
Mean AIC	4500.6	4544.4	4493.5	4544.6	1.019e4	1.019e4	1.023e4	1.014e4	1.640e4	1.628e4	1.625e4	1.633e4
Mean BIC	4505.0	4548.8	4497.9	4549.0	1.019e4	1.020e4	1.024e4	1.014e4	1.640e4	1.628e4	1.626e4	1.633e4
Marginal Cox												
Mean β	-0.692	-0.696	-0.713	-0.723	-0.696	-0.702	-0.700	-0.708	-0.695	-0.691	-0.694	-0.702
Mean SE	0.099	0.102	0.111	0.124	0.070	0.073	0.080	0.090	0.057	0.060	0.066	0.074
Empirical SE	0.104	0.102	0.118	0.135	0.073	0.078	0.082	0.091	0.058	0.063	0.062	0.079
Bias	0.001	-0.003	-0.020	-0.030	-0.003	-0.009	-0.007	-0.015	-0.002	0.002	-0.001	-0.009
MSE	0.010	0.010	0.013	0.016	0.005	0.005	0.006	0.008	0.003	0.004	0.004	0.005
95% CI coverage	0.948	0.940	0.930	0.924	0.932	0.938	0.942	0.950	0.940	0.942	0.964	0.942
Mean AIC	4500.6	4544.4	4493.5	4544.6	1.019e4	1.019e4	1.023e4	1.014e4	1.640e4	1.628e4	1.625e4	1.633e4
Mean BIC	4505.0	4548.8	4497.9	4549.0	1.019e4	1.020e4	1.024e4	1.014e4	1.640e4	1.628e4	1.626e4	1.633e4
Gamma Frailty												
Mean β	-0.695	-0.803	-0.987	-1.326	-0.697	-0.812	-0.972	-1.324	-0.697	-0.802	-0.966	-1.319
Mean SE	0.102	0.107	0.111	0.114	0.072	0.075	0.077	0.081	0.058	0.062	0.063	0.066
Empirical SE	0.104	0.106	0.131	0.148	0.073	0.082	0.089	0.108	0.058	0.068	0.070	0.088
Bias	-0.002	-0.109	-0.294	-0.633	-0.004	-0.119	-0.279	-0.630	-0.003	-0.109	-0.273	-0.626
MSE	0.010	0.023	0.098	0.413	0.005	0.020	0.084	0.404	0.003	0.016	0.079	0.396
95% CI coverage	0.958	0.944	0.906	0.874	0.940	0.920	0.912	0.860	0.944	0.922	0.918	0.834
Mean AIC	4499.3	4399.9	4167.7	3922.8	1.019e4	9899.8	9571.6	8862.4	1.639e4	1.583e4	1.525e4	1.439e4
Mean BIC	4508.1	4408.7	4176.5	3931.6	1.020e4	9910.0	9581.7	8872.6	1.641e4	1.583e4	1.527e4	1.440e4
Log Normal Frailty												
Mean β	-0.696	-0.802	-0.987	-1.327	-0.698	-0.812	-0.973	-1.326	-0.698	-0.801	-0.967	-1.321
Mean SE	0.102	0.107	0.111	0.114	0.072	0.075	0.077	0.081	0.058	0.062	0.063	0.065
Empirical SE	0.105	0.106	0.132	0.150	0.073	0.082	0.090	0.109	0.058	0.068	0.071	0.088
Bias	-0.003	-0.109	-0.294	-0.633	-0.005	-0.119	-0.279	-0.633	-0.004	-0.108	-0.274	-0.628
MSE	0.010	0.023	0.099	0.414	0.005	0.020	0.084	0.407	0.003	0.016	0.079	0.398
95% CI coverage	0.958	0.946	0.902	0.868	0.942	0.924	0.912	0.842	0.944	0.928	0.912	0.834
Mean AIC	4497.4	4399.3	4166.5	3921.1	1.018e4	9898.4	9568.5	8857.5	1.639e4	1.582e4	1.525e4	1.438e4
Mean BIC	4506.2	4408.1	4175.3	3930.0	1.019e4	9908.5	9578.7	8867.7	1.640e4	1.584e4	1.526e4	1.439e4

Table 1.2: Simulation results for models with 60% censoring (True model: marginal model)

	Clusters=30				Clusters=60				Clusters=90			
	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.50$	$\sigma = 0.75$	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.50$	$\sigma = 0.75$	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.50$	$\sigma = 0.75$
Naive Cox												
Mean β	-0.693	-0.700	-0.699	-0.713	-0.697	-0.687	-0.699	-0.703	-0.690	-0.688	-0.693	-0.711
Mean SE	0.135	0.135	0.135	0.137	0.095	0.095	0.095	0.095	0.077	0.077	0.078	0.078
Empirical SE	0.135	0.151	0.147	0.159	0.094	0.100	0.104	0.112	0.078	0.083	0.084	0.100
Bias	-<0.001	-0.007	-0.006	-0.020	-0.004	0.006	-0.006	-0.010	0.003	0.005	-<0.001	-0.018
MSE	0.018	0.018	0.018	0.019	0.009	0.009	0.009	0.009	0.006	0.006	0.006	0.006
95% CI coverage	0.948	0.916	0.922	0.902	0.938	0.934	0.926	0.904	0.954	0.946	0.944	0.884
Mean AIC	2553.6	2582.7	2588.6	2557.5	5767.2	5775.2	5823.8	5796.7	9247.9	9346.9	9215.0	9191.8
Mean BIC	2558.0	2587.1	2593.0	2561.9	5772.3	5780.3	5828.9	5801.8	9253.4	9352.4	9220.5	9197.3
Marginal Cox												
Mean β	-0.693	-0.700	-0.699	-0.713	-0.697	-0.687	-0.699	-0.703	-0.690	-0.688	-0.693	-0.711
Mean SE	0.131	0.134	0.142	0.156	0.094	0.097	0.102	0.113	0.076	0.079	0.084	0.092
Empirical SE	0.135	0.151	0.147	0.159	0.094	0.100	0.104	0.112	0.078	0.083	0.084	0.100
Bias	-<0.001	-0.007	-0.006	-0.020	-0.004	0.006	-0.006	-0.010	0.003	0.005	-<0.001	-0.018
MSE	0.017	0.018	0.020	0.025	0.009	0.009	0.010	0.013	0.006	0.006	0.007	0.009
95% CI coverage	0.942	0.914	0.936	0.944	0.934	0.940	0.940	0.948	0.948	0.952	0.960	0.934
Mean AIC	2553.6	2582.7	2588.6	2557.5	5767.2	5775.2	5823.8	5796.7	9247.9	9346.9	9215.0	9191.8
Mean BIC	2558.0	2587.1	2593.0	2561.9	5772.3	5780.3	5828.9	5801.8	9253.4	9352.4	9220.5	9197.3
Gamma Frailty												
Mean β	-0.696	-0.792	-0.929	-1.244	-0.699	-0.779	-0.932	-1.253	-0.691	-0.785	-0.927	-1.247
Mean SE	0.136	0.141	0.145	0.152	0.095	0.100	0.102	0.107	0.078	0.081	0.083	0.087
Empirical SE	0.136	0.157	0.161	0.186	0.095	0.100	0.110	0.134	0.078	0.086	0.091	0.116
Bias	-0.003	-0.099	-0.235	-0.550	-0.006	-0.085	-0.239	-0.559	0.002	-0.092	-0.233	-0.554
MSE	0.018	0.030	0.076	0.326	0.009	0.017	0.068	0.324	0.006	0.015	0.061	0.314
95% CI coverage	0.946	0.912	0.914	0.888	0.940	0.954	0.926	0.886	0.948	0.932	0.940	0.856
Mean AIC	2552.0	2471.3	2338.7	2105.2	5763.8	5537.4	5315.5	4854.6	9244.1	8982.9	8454.7	7801.0
Mean BIC	2560.8	2480.1	2347.5	2114.0	5774.0	5547.6	5325.7	4864.8	9255.1	8993.9	8465.7	7812.0
Log Normal Frailty												
Mean β	-0.697	-0.793	-0.931	-1.247	-0.7001	-0.779	-0.935	-1.255	-0.692	-0.786	-0.929	-1.250
Mean SE	0.136	0.141	0.145	0.152	0.096	0.100	0.102	0.106	0.078	0.081	0.083	0.087
Empirical SE	0.136	0.158	0.164	0.190	0.095	0.100	0.111	0.135	0.078	0.087	0.093	0.118
Bias	-0.004	-0.099	-0.238	-0.553	-0.007	-0.086	-0.242	-0.562	0.001	-0.092	-0.236	-0.557
MSE	0.018	0.030	0.078	0.330	0.009	0.017	0.069	0.327	0.006	0.015	0.063	0.318
95% CI coverage	0.948	0.908	0.916	0.884	0.940	0.956	0.932	0.880	0.948	0.932	0.934	0.846
Mean AIC	2550.1	2471.3	2339.0	2105.9	5760.6	5537.5	5315.8	4855.3	9239.6	8983.1	8455.3	7800.8
Mean BIC	2558.9	2480.1	2347.8	2114.7	5770.8	5547.7	5326.0	4865.5	9250.6	8994.1	8466.3	7811.8

models with lower censoring. This inflation is not as pronounced when the number of clusters is higher.

When there is no dependence within center ($\sigma = 0$), the conditional coefficient and the marginal coefficient should be the same. Under this constraint, the frailty models perform well (small bias) and have similar performance to the marginal model. When the correlation within center increases, the estimate of the treatment effect becomes larger in magnitude (becomes more negative). This leads to a larger difference between the conditional and marginal coefficients. This result is expected theoretically (Henderson & Oman, 1999) and can be seen from the relationship between the conditional and marginal parameters in the positive stable model (1.6). For the frailty models, the empirical coverage of the 95% confidence intervals tends to decrease with increasing correlation. This may be a consequence of the standard errors being estimated under the assumption that θ is a fixed, rather than estimated, quantity (Therneau & Grambsch, 2000). It also appears that the model-based standard errors tend to be too small for larger correlations as seen by the larger empirical standard errors.

The gamma and log normal frailty models give very similar results for the estimates of both the treatment effect and the standard errors. This leads to bias and MSE estimates that are very close. The AIC and BIC are also quite similar between the two frailty models and are always higher for the population-averaged models than for the frailty models. As was the case for the population-averaged models, standard errors are inflated for the settings with higher censoring. The coefficients also tend to be larger (closer to 0) under the higher censoring settings.

The relationship between the marginal and conditional coefficients can be seen visually by looking at Figure 1.1, which is representative of the different simulation settings. When there is no dependence within center, the marginal and conditional coefficients are quite close. However, as the amount of dependence

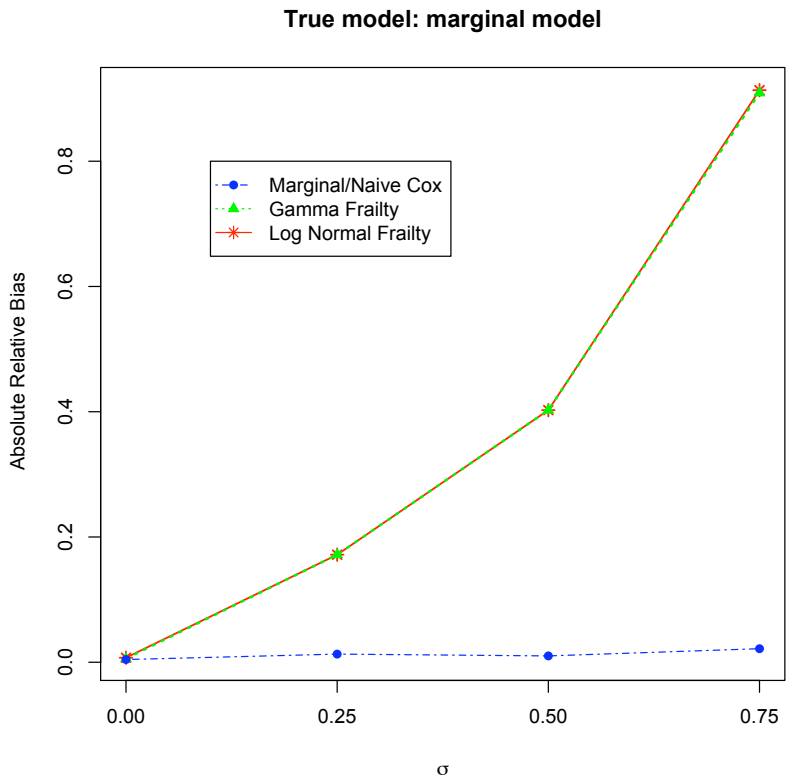

Figure 1.1: Plots of absolute relative bias versus σ for 60 clusters and 30% censoring where the true model is the marginal model.

increases, the difference between the conditional and true marginal coefficients increases. In fact, for large amounts of dependence within center, we see that the difference between the two types of estimates is quite drastic. From the picture, it is also clear that the gamma and log normal frailty models perform very similarly.

1.4.2 True Model: Conditional Model

In multicenter clinical trials we may also be interested in center effects or a center-by-treatment interaction. In this case, the use of a frailty model would be reasonable. In order to evaluate and compare the performance of the gamma frailty model and the log normal frailty model, data were generated to follow a pro-

portional hazards model with a known conditional treatment effect. As in the marginal setting, we consider estimates for the treatment effect (β), the model-based standard error, the empirical standard error, the bias, the mean squared error, the empirical coverage of 95 percent confidence intervals, AIC and BIC. We can fit the population-averaged models to the data as well but the bias (and by extension the MSE) represents the difference from the true conditional effect, not the difference from the true marginal effect. These simulations, therefore, do not let us see how well models estimate the true marginal effect. The simulations do, however, allow us to consider relationships between marginal and conditional coefficients across different settings.

Method of Simulation

The clustered failure time data with a known conditional treatment effect were simulated using the approach of Bender et al. (2005), with the addition of a frailty term. Exponential clustered survival times (T_{ij}) can be simulated using the following equation:

$$T_{ij} = -\frac{\log(U_{ij})}{\lambda \exp(\beta^T Z_{ij} + \gamma_j)} \tag{1.9}$$

where U_{ij} is generated from a Uniform(0,1) distribution and λ is the scale parameter for the exponential distribution (chosen to be 2 for our simulations). γ_j in (1.9) is the frailty term (shared by all patients in the same center) that controls the dependence within each cluster. For our simulations, the γ_js were generated either from the log of a Gamma $(1/\theta, 1/\theta)$ distribution (true model: gamma frailty model) or from a Normal(0, θ) distribution (true model: log normal frailty model).

As in the marginal setting, twenty-four different simulation settings were considered for both the gamma frailty and log normal frailty settings: 2 levels of censoring (30%, 60%), 4 levels of dependence within center (θ= 0, 0.67, 2, 6), and 3 levels of cluster number (30, 60, 90). The values of θ (variance of the frailty term)

were chosen to correspond to Kendall's τ values of 0, 0.25, 0.5, and 0.75 under the gamma frailty model. For each setting, 500 simulations were carried out and a single treatment covariate was generated from a Bernoulli(0.5) distribution. The conditional treatment effect (β) was set to log(0.5) = -0.693 for each setting and cluster size was randomly generated to allow clusters of varying sizes between 10 and 50 with a mean cluster size of 30.

Results

Table 1.3 presents simulation results for the models with 30% censoring where the true model is the gamma frailty model and Table 1.4 presents similar results for models with 60% censoring. We will first focus on the performance of the two frailty models. There are small biases and the empirical coverage of the confidence intervals are close to 95% in all cases, suggesting that the gamma frailty model performs well in all the settings. This is to be expected since the true frailty is from a gamma distribution. However, the log normal frailty model also performs very well and even outperforms the gamma frailty model in some cases (smaller bias). For both models, when the number of clusters is smaller, there is some tendency for the frailty model to underestimate the standard errors, particularly with higher correlation within center. Higher censoring has the effect of increasing the standard errors. When there is dependence within center ($\theta > 0$), the AIC and BIC do tend to be lower for the gamma frailty model as compared to the log normal frailty model. This difference is most apparent for the higher correlations. When there is no dependence within center ($\theta = 0$), the AIC and BIC are lower for the log normal frailty model than the gamma frailty model.

When there is no dependence within center ($\theta = 0$), the conditional coefficient and the marginal coefficient should be the same. In this setting, both the naive Cox model and the marginal Cox model perform similarly to the frailty

Table 1.3: Simulation results for models with 30% censoring (True model: gamma frailty model)

	Clusters=30				Clusters=60				Clusters=90			
	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$
Naive Cox												
Mean β	-0.689	-0.475	-0.321	-0.217	-0.691	-0.462	-0.313	-0.211	-0.697	-0.454	-0.309	-0.217
Mean SE	0.089	0.090	0.098	0.123	0.063	0.063	0.069	0.087	0.051	0.052	0.056	0.070
Empirical SE	0.093	0.105	0.102	0.124	0.062	0.068	0.072	0.088	0.054	0.058	0.059	0.073
Bias	0.004	0.219	0.373	0.477	0.002	0.231	0.380	0.482	-0.004	0.239	0.384	0.476
MSE	0.008	0.056	0.149	0.242	0.004	0.058	0.149	0.240	0.003	0.060	0.150	0.231
95% CI coverage	0.948	0.896	0.942	0.950	0.954	0.940	0.948	0.948	0.946	0.912	0.954	0.942
Mean AIC	5938.2	5592.5	4801.5	3255.6	1.347e4	1.273e4	1.084e4	7201.2	2.149e4	2.025e4	1.729e4	1.153e4
Mean BIC	5942.6	5596.9	4805.9	3260.0	1.348e4	1.273e4	1.085e4	7206.3	2.150e4	2.026e4	1.729e4	1.154e4
Marginal Cox												
Mean β	-0.689	-0.475	-0.321	-0.217	-0.691	-0.462	-0.313	-0.211	-0.697	-0.454	-0.309	-0.217
Mean SE	0.088	0.092	0.099	0.122	0.062	0.067	0.071	0.087	0.051	0.055	0.059	0.072
Empirical SE	0.093	0.105	0.102	0.124	0.062	0.068	0.072	0.088	0.054	0.058	0.059	0.073
Bias	0.004	0.219	0.373	0.477	0.002	0.231	0.380	0.482	-0.004	0.239	0.384	0.476
MSE	0.008	0.056	0.149	0.242	0.004	0.058	0.149	0.240	0.003	0.060	0.151	0.231
95% CI coverage	0.940	0.902	0.942	0.950	0.952	0.948	0.958	0.952	0.940	0.930	0.964	0.946
Mean AIC	5938.2	5592.5	4801.5	3255.6	1.347e4	1.273e4	1.084e4	7201.2	2.149e4	2.025e4	1.729e4	1.153e4
Mean BIC	5942.6	5596.9	4805.9	3260.0	1.348e4	1.273e4	1.085e4	7206.3	2.150e4	2.026e4	1.729e4	1.154e4
Gamma Frailty												
Mean β	-0.691	-0.709	-0.693	-0.691	-0.692	-0.701	-0.694	-0.693	-0.698	-0.700	-0.693	-0.696
Mean SE	0.090	0.096	0.106	0.132	0.063	0.068	0.074	0.093	0.052	0.055	0.060	0.075
Empirical SE	0.092	0.098	0.106	0.137	0.063	0.070	0.072	0.099	0.054	0.056	0.064	0.076
Bias	0.002	-0.016	<0.001	0.002	0.001	-0.008	-0.001	<0.001	-0.005	-0.006	<0.001	-0.003
MSE	0.008	0.010	0.011	0.018	0.004	0.005	0.005	0.008	0.003	0.003	0.004	0.006
95% CI coverage	0.954	0.946	0.952	0.932	0.954	0.940	0.952	0.926	0.944	0.942	0.926	0.934
Mean AIC	5934.9	5272.3	4181.5	2466.6	1.347e4	1.207e4	9565.9	5591.2	2.149e4	1.926e4	1.536e4	9102.7
Mean BIC	5939.3	5276.7	4185.9	2471.0	1.347e4	1.207e4	9571.0	5596.3	2.149e4	1.927e4	1.537e4	9108.2
Log Normal Frailty												
Mean β	-0.693	-0.704	-0.692	-0.687	-0.693	-0.696	-0.693	-0.689	-0.699	-0.694	-0.692	-0.693
Mean SE	0.090	0.096	0.106	0.132	0.063	0.067	0.074	0.093	0.052	0.055	0.060	0.075
Empirical SE	0.092	0.098	0.106	0.136	0.063	0.070	0.072	0.099	0.054	0.055	0.063	0.076
Bias	<0.001	-0.011	0.001	0.006	-<0.001	-0.003	<0.001	0.004	-0.006	-0.001	0.001	<0.001
MSE	0.008	0.009	0.011	0.018	0.004	0.005	0.005	0.009	0.003	0.003	0.004	0.006
95% CI coverage	0.954	0.940	0.952	0.932	0.952	0.944	0.952	0.930	0.942	0.942	0.926	0.938
Mean AIC	5932.7	5273.4	4183.5	2470.6	1.347e4	1.207e4	9570.0	5599.0	2.148e4	1.926e4	1.537e4	9114.4
Mean BIC	5937.1	5277.8	4187.9	2474.9	1.347e4	1.208e4	9575.1	5604.1	2.149e4	1.927e4	1.538e4	9119.9

Table 1.4: Simulation results for models with 60% censoring (True model: gamma frailty model)

	Clusters=30				Clusters=60				Clusters=90			
	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$
Naive Cox												
Mean β	-0.688	-0.541	-0.411	-0.304	-0.694	-0.528	-0.406	-0.308	-0.696	-0.522	-0.409	-0.294
Mean SE	0.108	0.114	0.127	0.160	0.076	0.080	0.090	0.112	0.062	0.065	0.073	0.090
Empirical SE	0.109	0.121	0.139	0.164	0.076	0.085	0.090	0.116	0.064	0.065	0.075	0.094
Bias	0.005	0.152	0.282	0.389	-0.001	0.165	0.287	0.385	-0.003	0.171	0.285	0.399
MSE	0.012	0.036	0.096	0.177	0.006	0.034	0.090	0.161	0.004	0.033	0.086	0.167
95% CI coverage	0.958	0.934	0.922	0.940	0.950	0.948	0.940	0.938	0.942	0.958	0.938	0.936
Mean AIC	4023.1	3541.7	2943.1	1967.2	9074.2	7972.1	6548.6	4365.8	1.447e4	1.285e4	1.052e4	7052.9
Mean BIC	4027.5	3546.1	2947.5	1971.6	9079.3	7977.2	6553.7	4370.9	1.448e4	1.286e4	1.052e4	7058.4
Marginal Cox												
Mean β	-0.688	-0.541	-0.411	-0.304	-0.694	-0.528	-0.406	-0.308	-0.696	-0.522	-0.409	-0.294
Mean SE	0.104	0.113	0.126	0.157	0.074	0.080	0.091	0.113	0.061	0.066	0.075	0.092
Empirical SE	0.109	0.121	0.139	0.164	0.076	0.085	0.090	0.116	0.064	0.065	0.075	0.094
Bias	0.005	0.152	0.282	0.389	-0.001	0.165	0.287	0.385	-0.003	0.171	0.285	0.399
MSE	0.011	0.036	0.095	0.176	0.006	0.034	0.091	0.161	0.004	0.034	0.087	0.168
95% CI coverage	0.944	0.928	0.916	0.932	0.948	0.946	0.944	0.944	0.938	0.960	0.946	0.940
Mean AIC	4023.1	3541.7	2943.1	1967.2	9074.2	7972.1	6548.6	4365.8	1.447e4	1.285e4	1.052e4	7052.9
Mean BIC	4027.5	3546.1	2947.5	1971.6	9079.3	7977.2	6553.7	4370.9	1.448e4	1.286e4	1.052e4	7058.4
Gamma Frailty												
Mean β	-0.690	-0.697	-0.693	-0.686	-0.696	-0.698	-0.691	-0.696	-0.698	-0.694	-0.694	-0.690
Mean SE	0.108	0.121	0.136	0.172	0.076	0.085	0.096	0.120	0.062	0.069	0.077	0.097
Empirical SE	0.109	0.126	0.140	0.180	0.076	0.088	0.096	0.122	0.064	0.067	0.076	0.095
Bias	0.003	-0.004	$<$0.001	0.007	-0.002	-0.005	0.002	-0.003	-0.005	-0.001	-0.001	0.004
MSE	0.012	0.015	0.018	0.030	0.006	0.007	0.009	0.014	0.004	0.005	0.006	0.009
95% CI coverage	0.956	0.938	0.932	0.940	0.950	0.942	0.952	0.940	0.942	0.962	0.956	0.942
Mean AIC	4020.0	3334.9	2558.7	1483.1	9069.6	7541.1	5763.4	3377.2	1.447e4	1.220e4	9321.2	5549.7
Mean BIC	4024.4	3339.3	2563.1	1487.5	9074.7	7546.2	5768.5	3382.3	1.447e4	1.221e4	9326.7	5555.2
Log Normal Frailty												
Mean β	-0.691	-0.694	-0.692	-0.681	-0.697	-0.695	-0.690	-0.691	-0.699	-0.691	-0.693	-0.685
Mean SE	0.108	0.121	0.136	0.172	0.076	0.085	0.096	0.120	0.062	0.069	0.077	0.097
Empirical SE	0.110	0.125	0.140	0.178	0.076	0.087	0.095	0.121	0.064	0.067	0.076	0.094
Bias	0.002	-0.001	0.001	0.012	-0.003	-0.002	0.003	0.002	-0.006	0.002	$<$0.001	0.009
MSE	0.012	0.015	0.018	0.030	0.006	0.007	0.009	0.014	0.004	0.005	0.006	0.009
95% CI coverage	0.952	0.938	0.938	0.942	0.950	0.938	0.954	0.938	0.942	0.960	0.956	0.944
Mean AIC	4018.2	3336.7	2562.2	1488.0	9066.5	7544.8	5770.5	3387.1	1.446e4	1.221e4	9331.9	5564.5
Mean BIC	4022.6	3341.1	2566.6	1492.4	9071.6	7549.9	5775.6	3392.2	1.447e4	1.221e4	9337.4	5570.0

model. However, when the correlation within center increases, the estimate of the treatment effect for the population-averaged models becomes smaller in magnitude (becomes less negative). This leads to a larger difference between the conditional and marginal coefficients. As expected, the bias is the same for the naive Cox model and the marginal Cox model. Interestingly, however, the standard error estimates are similar for the two models as well. This suggests that when the true model is a gamma frailty model, performance does not differ much between the naive Cox and marginal Cox models even though dependence within center is present. AIC and BIC are always higher for the population-averaged models than for the frailty models.

Table 1.5 presents simulation results for the models with 30% censoring where the true model is the log normal frailty model and Table 1.6 presents similar results for models with 60% censoring. When the true model is the log normal frailty model, the log normal frailty model performs well in all settings as seen by small biases and empirical coverage of confidence intervals that are close to 95%. The gamma frailty model also performs very well and again even outperforms the log normal frailty model in some cases (smaller bias). As in the gamma case, when the number of clusters is smaller, there is some tendency for the frailty model to underestimate the standard errors, particularly with higher correlation within center and higher censoring has the effect of increasing the standard errors. When there is dependence within center ($\theta > 0$), the AIC and BIC tend to be very similar for the two frailty models. Even though the true model is the log normal frailty model, in many of the settings, the AIC and BIC are lower for the gamma frailty model.

When there is no dependence within center ($\theta = 0$), the naive Cox model and the marginal Cox model again perform similarly to the frailty models. When the correlation within center increases, results are similar as in the gamma case. The estimate of the treatment effect for the population-averaged models becomes

Table 1.5: Simulation results for models with 30% censoring (True model: log normal frailty model)

	Clusters=30				Clusters=60				Clusters=90			
	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$
Naive Cox												
Mean β	-0.706	-0.529	-0.403	-0.292	-0.697	-0.525	-0.407	-0.285	-0.690	-0.521	-0.404	-0.283
Mean SE	0.099	0.099	0.100	0.103	0.070	0.069	0.070	0.073	0.057	0.057	0.058	0.059
Empirical SE	0.099	0.099	0.112	0.106	0.069	0.072	0.076	0.072	0.057	0.063	0.060	0.066
Bias	-0.012	0.164	0.290	0.401	-0.004	0.168	0.286	0.408	0.003	0.173	0.289	0.410
MSE	0.010	0.037	0.094	0.172	0.005	0.033	0.087	0.172	0.003	0.033	0.087	0.172
95% CI coverage	0.944	0.950	0.940	0.954	0.958	0.948	0.934	0.944	0.948	0.922	0.952	0.922
Mean AIC	4788.4	4720.9	4584.4	4406.9	1.070e4	1.059e4	1.035e4	9834.1	1.724e4	1.684e4	1.638e4	1.571e4
Mean BIC	4792.8	4725.3	4588.8	4411.3	1.070e4	1.060e4	1.036e4	9839.2	1.724e4	1.685e4	1.639e4	1.572e4
Marginal Cox												
Mean β	-0.706	-0.529	-0.403	-0.292	-0.697	-0.525	-0.407	-0.285	-0.690	-0.521	-0.404	-0.283
Mean SE	0.096	0.097	0.101	0.105	0.069	0.071	0.073	0.076	0.057	0.059	0.060	0.062
Empirical SE	0.099	0.099	0.112	0.106	0.069	0.072	0.076	0.072	0.057	0.063	0.060	0.066
Bias	-0.012	0.164	0.290	0.401	-0.004	0.168	0.286	0.408	0.003	0.173	0.289	0.410
MSE	0.009	0.036	0.094	0.172	0.005	0.033	0.087	0.172	0.003	0.033	0.087	0.172
95% CI coverage	0.930	0.946	0.942	0.960	0.958	0.950	0.940	0.948	0.948	0.926	0.956	0.930
Mean AIC	4788.4	4720.9	4584.4	4406.9	1.070e4	1.059e4	1.035e4	9834.1	1.724e4	1.684e4	1.638e4	1.571e4
Mean BIC	4792.8	4725.3	4588.8	4411.3	1.070e4	1.060e4	1.036e4	9839.2	1.724e4	1.685e4	1.639e4	1.572e4
Gamma Frailty												
Mean β	-0.708	-0.697	-0.685	-0.697	-0.698	-0.698	-0.688	-0.690	-0.691	-0.696	-0.693	-0.688
Mean SE	0.099	0.104	0.107	0.110	0.070	0.074	0.075	0.078	0.057	0.060	0.061	0.063
Empirical SE	0.099	0.100	0.111	0.114	0.069	0.072	0.079	0.080	0.057	0.066	0.059	0.065
Bias	-0.014	-0.003	0.008	-0.003	-0.005	-0.005	0.006	0.004	0.002	-0.003	0.001	0.005
MSE	0.010	0.011	0.012	0.012	0.005	0.005	0.006	0.006	0.003	0.004	0.004	0.004
95% CI coverage	0.948	0.968	0.946	0.948	0.958	0.956	0.924	0.938	0.950	0.930	0.956	0.934
Mean AIC	4784.8	4493.1	4133.6	3634.2	1.070e4	1.012e4	9424.0	8254.5	1.723e4	1.614e4	1.498e4	1.331e4
Mean BIC	4789.3	4497.5	4138.0	3638.6	1.070e4	1.013e4	9429.1	8259.6	1.724e4	1.615e4	1.499e4	1.331e4
Log Normal Frailty												
Mean β	-0.709	-0.695	-0.687	-0.699	-0.699	-0.696	-0.689	-0.693	-0.692	-0.694	-0.695	-0.692
Mean SE	0.099	0.104	0.107	0.110	0.070	0.074	0.075	0.078	0.057	0.060	0.061	0.063
Empirical SE	0.099	0.100	0.111	0.114	0.070	0.072	0.080	0.080	0.057	0.066	0.059	0.065
Bias	-0.016	-0.002	0.006	-0.005	-0.006	-0.002	0.004	$<$0.001	0.001	-$<$0.001	-0.002	0.001
MSE	0.010	0.011	0.011	0.012	0.005	0.005	0.006	0.006	0.003	0.004	0.004	0.004
95% CI coverage	0.948	0.968	0.950	0.948	0.958	0.956	0.926	0.940	0.946	0.930	0.956	0.936
Mean AIC	4783.1	4493.3	4133.3	3634.6	1.069e4	1.012e4	9422.6	8254.1	1.722e4	1.614e4	1.498e4	1.331e4
Mean BIC	4787.5	4497.7	4137.7	3639.0	1.070e4	1.013e4	9427.7	8259.2	1.723e4	1.615e4	1.499e4	1.331e4

Table 1.6: Simulation results for models with 60% censoring (True model: log normal frailty model)

	Clusters=30				Clusters=60				Clusters=90			
	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$	$\theta = 0$	$\theta = 0.67$	$\theta = 2$	$\theta = 6$
Naive Cox												
Mean β	-0.696	-0.566	-0.451	-0.322	-0.693	-0.569	-0.447	-0.313	-0.692	-0.567	-0.440	-0.313
Mean SE	0.130	0.126	0.125	0.123	0.091	0.089	0.088	0.086	0.074	0.073	0.071	0.070
Empirical SE	0.137	0.129	0.128	0.131	0.094	0.095	0.088	0.090	0.072	0.075	0.075	0.071
Bias	-0.003	0.127	0.242	0.371	<0.001	0.124	0.247	0.380	0.001	0.126	0.253	0.380
MSE	0.017	0.032	0.074	0.153	0.008	0.023	0.069	0.152	0.006	0.021	0.069	0.149
95% CI coverage	0.932	0.948	0.946	0.932	0.956	0.928	0.952	0.934	0.956	0.952	0.948	0.962
Mean AIC	2783.8	2940.1	3027.8	3185.6	6265.7	6569.8	6841.6	7144.6	9982.1	1.045e4	1.089e4	1.149e4
Mean BIC	2788.2	2944.5	3032.2	3190.0	6270.8	6574.9	6846.7	7149.7	9987.6	1.046e4	1.089e4	1.149e4
Marginal Cox												
Mean β	-0.696	-0.566	-0.451	-0.322	-0.693	-0.569	-0.447	-0.313	-0.692	-0.567	-0.440	-0.313
Mean SE	0.126	0.124	0.126	0.125	0.089	0.090	0.090	0.090	0.073	0.073	0.074	0.073
Empirical SE	0.137	0.129	0.128	0.131	0.094	0.095	0.088	0.090	0.072	0.075	0.075	0.071
Bias	-0.003	0.127	0.242	0.371	<0.001	0.124	0.247	0.380	0.001	0.126	0.253	0.380
MSE	0.016	0.032	0.074	0.153	0.008	0.023	0.069	0.152	0.005	0.021	0.070	0.150
95% CI coverage	0.928	0.940	0.946	0.932	0.948	0.932	0.962	0.942	0.956	0.952	0.954	0.968
Mean AIC	2783.8	2940.1	3027.8	3185.6	6265.7	6569.8	6841.6	7144.6	9982.1	1.045e4	1.089e4	1.149e4
Mean BIC	2788.2	2944.5	3032.2	3190.0	6270.8	6574.9	6846.7	7149.7	9987.6	1.046e4	1.089e4	1.149e4
Gamma Frailty												
Mean β	-0.699	-0.690	-0.687	-0.698	-0.695	-0.696	-0.687	-0.685	-0.693	-0.692	-0.682	-0.688
Mean SE	0.130	0.132	0.133	0.132	0.091	0.093	0.093	0.092	0.075	0.076	0.076	0.074
Empirical SE	0.138	0.132	0.135	0.138	0.095	0.101	0.092	0.092	0.072	0.076	0.075	0.073
Bias	-0.006	0.003	0.006	-0.004	-0.002	-0.003	0.006	0.008	<0.001	0.001	0.012	0.006
MSE	0.017	0.018	0.018	0.017	0.008	0.009	0.009	0.009	0.006	0.006	0.006	0.006
95% CI coverage	0.934	0.956	0.956	0.936	0.956	0.938	0.948	0.966	0.956	0.948	0.950	0.956
Mean AIC	2780.4	2781.4	2693.7	2569.2	6260.1	6247.3	6151.7	5879.7	9978.0	9966.2	9836.5	9562.9
Mean BIC	2784.8	2785.8	2698.1	2573.6	6265.2	6252.4	6156.8	5884.8	9983.5	9971.7	9842.0	9568.4
Log Normal Frailty												
Mean β	-0.700	-0.691	-0.691	-0.701	-0.696	-0.696	-0.691	-0.689	-0.694	-0.693	-0.685	-0.692
Mean SE	0.130	0.132	0.133	0.132	0.092	0.093	0.093	0.092	0.075	0.076	0.076	0.074
Empirical SE	0.139	0.132	0.135	0.139	0.095	0.100	0.093	0.092	0.072	0.076	0.075	0.074
Bias	-0.007	0.002	0.002	-0.008	-0.003	-0.003	0.002	0.004	-<0.001	0.001	0.008	0.001
MSE	0.017	0.018	0.018	0.017	0.008	0.009	0.009	0.008	0.006	0.006	0.006	0.006
95% CI coverage	0.934	0.956	0.954	0.940	0.956	0.942	0.942	0.966	0.956	0.948	0.948	0.954
Mean AIC	2778.7	2781.8	2694.5	2570.9	6257.0	6248.3	6153.5	5882.3	9973.7	9967.8	9838.6	9565.5
Mean BIC	2783.1	2786.2	2698.9	2575.3	6262.1	6253.4	6158.6	5887.4	9979.2	9973.3	9844.1	9571.0

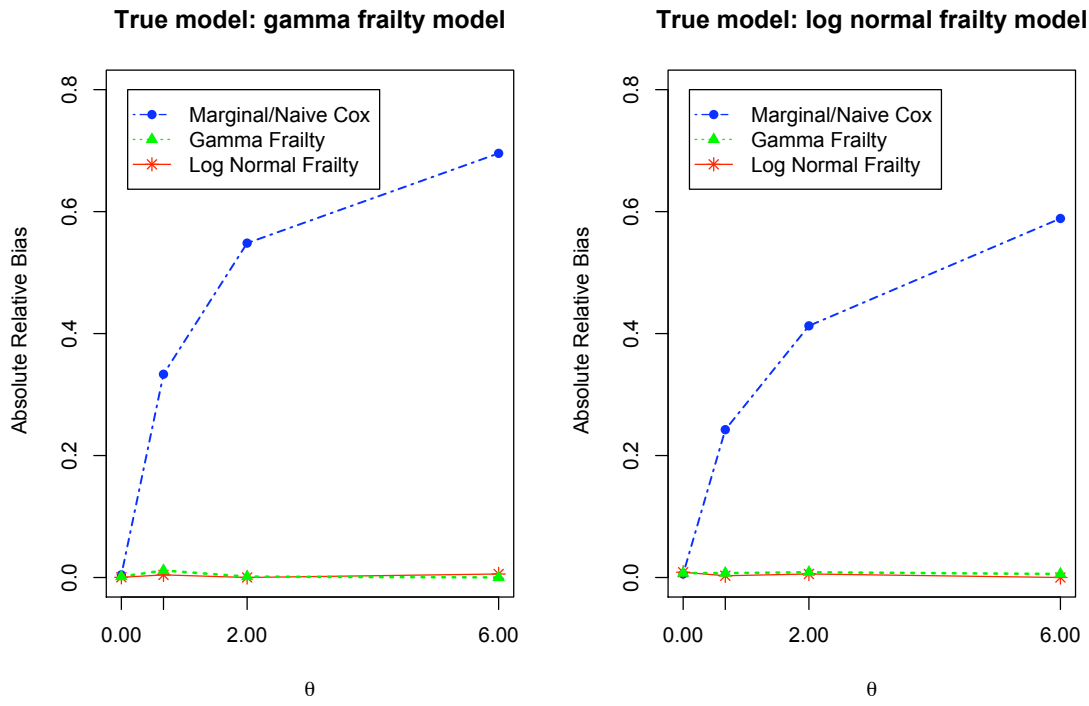

Figure 1.2: Plots of absolute relative bias versus θ for 60 clusters and 30% censoring where the true model is a conditional model.

smaller in magnitude (becomes less negative). Results suggest that when the true model is a log normal frailty model, performance does not differ much between the naive Cox and marginal Cox models even though dependence within center is present.

The relationship between the marginal and conditional coefficients can be seen visually by looking at Figure 1.2. When there is no dependence within center, the marginal and conditional coefficients are quite close. However, as the amount of dependence increases, the difference between the marginal and true conditional coefficient increases. The pattern of the absolute relative bias is similar for both the true gamma model and the true log normal model. However, the bias is somewhat smaller in the case of the log normal frailty model. From the picture, it is also clear that the gamma and log normal frailty models perform very similarly.

1.5 Data Example

The Children's Oncology Group (COG) study 1961 was used to study postinduction intensification (PII) in children and adolescents with"higher risk" acute lymphoblastic leukemia (ALL) (Seibel et al., 2008). The goal of the study was to see if longer duration PII and increased strength PII would improve the survival for children with"higher risk" ALL and a rapid marrow response to induction therapy. In order to ascertain the relative benefit or harm of each variation of PII, a 2X2 factorial design was used. For the purpose of this analysis two patient populations were studied: the full sample of 1299 patients, and a subset of 460 patients with enlarged livers, which represents a subgroup of sicker patients. The endpoint of interest in this case is overall survival.

The four methods focused on in the paper were applied to both the full dataset and the enlarged liver subset. There was no evidence of a duration by strength interaction in either case. Therefore, the various models include the two treatment variables, increased strength and duration, and three other baseline characteristics: gender, age, and platelet count at diagnosis. Using the full dataset, very little clustering is seen as evidenced by the Kendall's τ estimate from the gamma frailty model of < 0.001 (95% CI: (-0.015, 0.015)). On the other hand, there is a suggestion of a clustering effect for the enlarged liver subset as seen by the Kendall's τ estimate of 0.192 for the gamma frailty model. However, the 95% confidence interval of (-0.110, 0.494) suggests this effect might not be significant. 95% confidence intervals for Kendall's τ were calculated using the jackknife leaving out one cluster at a time. Table 1.7 presents data analysis results for the full dataset and Table 1.8 presents results for the enlarged liver subset.

In the full dataset, we can see that there is a significant effect of increased strength, but a quite insignificant effect of increased duration. Results in this case suggest that increased strength is associated with improved survival for the overall

Table 1.7: COG data analysis results - full dataset

Parameter	Estimate	SE	p-value
Naive Cox			
Increased Strength	-0.436	0.150	0.004
Increased Duration	-0.108	0.147	0.460
AIC = 2594.4; BIC = 2630.6			
Marginal Cox			
Increased Strength	-0.436	0.141	0.002
Increased Duration	-0.108	0.145	0.456
AIC = 2594.4; BIC = 2630.6			
Gamma Frailty			
Increased Strength	-0.436	0.150	0.004
Increased Duration	-0.108	0.147	0.460
AIC = 2596.4; BIC = 2637.7			
Log Normal Frailty			
Increased Strength	-0.438	0.150	0.003
Increased Duration	-0.107	0.147	0.470
AIC = 2592.7.4; BIC = 2634.0			

*Models also include gender, age, and platelet count at diagnosis

Table 1.8: COG data analysis results - enlarged liver subset

Parameter	Estimate	SE	p-value
Naive Cox			
Increased Strength	-0.429	0.247	0.082
Increased Duration	0.409	0.250	0.102
	AIC = 809.7; BIC = 838.6		
Marginal Cox			
Increased Strength	-0.429	0.229	0.061
Increased Duration	0.409	0.242	0.091
	AIC = 809.7; BIC = 838.6		
Gamma Frailty			
Increased Strength	-0.466	0.253	0.065
Increased Duration	0.476	0.257	0.064
	AIC = 764.5; BIC = 797.5		
Log Normal Frailty			
Increased Strength	-0.462	0.253	0.067
Increased Duration	0.488	0.257	0.057
	AIC = 763.6; BIC = 796.7		

*Models also include gender, age, and platelet count at diagnosis

study population. Also of particular interest in this case is the similarity between the coefficient estimates for all of the models. Even though some of the parameters have conditional interpretations and others have marginal interpretations, in this case there is very little difference between them. This is to be expected, however, since the amount of dependence is so small. We also note that the gamma and log normal frailty models give almost identical results, suggesting that the choice of frailty is not crucial. In this setting we would recommend using the traditional Cox model for analysis since clustering within center is not an issue.

In the enlarged liver subset, we see that there is a marginally significant effect of both increased strength and increased duration for most of the models. Results suggest that increased strength is associated with improved survival and that increased duration is associated with worse survival. Unlike the full dataset, in this subset, there are differences between the marginal and conditional coefficient estimates. We can see that the marginal estimates are closer to 0 than the conditional estimates. This illustrates that even with a minimal amount of dependence in the dataset, the differences between marginal and conditional coefficients are apparent. In this subset, we would recommend using the marginal Cox model for analysis since there is some evidence of clustering within center and since the aim of the trial is to determine the effectiveness of treatment at the population level.

1.6 Discussion

This paper has reviewed current methods for dealing with clustered failure time data that may arise from multicenter clinical trials. The simulation study has shown that when your data are clustered and marginally follow a proportional hazards model, the marginal Cox model is readily able to estimate a population-averaged treatment effect in a variety of settings. The naive Cox model does a reasonable job when the within cluster dependence is small, but should not be used

when the clustering effect is more significant. When your data are clustered and follow a proportional hazards model with a frailty term, both the gamma frailty model and the log normal frailty model do a good job of estimating the conditional treatment effect in a variety of settings. Both models perform well whether the true frailty is gamma or log normal.

The population-averaged models perform well when the true model is a marginal model and the frailty models perform well when the true model is a conditional model. However, when a marginal model is fit to a true conditional model or a frailty model is fit to a true marginal model, large differences in the coefficients can exist. When the correlation within center is small, there is not a big difference between the marginal and conditional coefficients. A large amount of dependence within center, however, means a large difference between the two types of coefficients. This implies that a correct interpretation of the chosen model is crucial.

Because the interpretation does differ between the frailty models and the naive Cox and marginal Cox models, it would be ideal if there was some way of using the data to perform model selection. Choosing the model with the lowest AIC or BIC would be a logical choice for doing this model selection. However, it turns out that this is not a good approach. The AIC and BIC are identical for the naive Cox and marginal Cox models because they are both based on the same partial likelihood. This means that AIC or BIC cannot be used to choose between these two models. The marginal model takes clustering into account in calculating the standard error. However, the marginal model does not adjust the likelihood in any way, so the AIC/BIC for the marginal Cox model does not account for clustering. On the other hand, the likelihoods for the frailty models do account for clustering. Therefore, as the simulations show, the AIC/BIC for the frailty models is smaller than the AIC/BIC for the marginal Cox model even when the true model is a marginal model and the frailty model leads to biased estimates. This means

that using AIC or BIC is not a reliable way to choose between a frailty model and the marginal Cox model when dependence exists within center. It may be possible to use AIC or BIC to choose between frailty models. However, even this seems somewhat unreliable since in some simulation settings the AIC/BIC was smaller for the gamma model than the log normal model even when the true model was the log normal frailty model.

A possible alternative for model selection would be the use of some testing procedures. For example, a rejection of the proportional hazards assumption could be seen as evidence in favor of most frailty models (except for the positive stable frailty model). However, tests for proportional hazards often have limited power (Lin & Wei, 1991; Grambsch & Therneau, 1994), especially when the sample size for a study is moderate. In addition, non-proportionality could be due to an incorrect functional form of a covariate or missing covariates (Therneau & Grambsch, 2000). Score tests such as the ones proposed by Gray (1995) and Commenges & Andersen (1995) could also be an option, but it is unclear that these tests would help with choosing between the marginal Cox model and a frailty model when clustering truly exists. Due to the uncertainty of what conclusions can be reliably drawn from such tests, we opt not to consider testing here for model selection.

Since AIC and BIC are not useful for model selection and testing procedures may not be reliable, there need to be some guidelines about how to choose between the models. We recommend choosing between marginal and conditional models based on the scientific question of interest. A marginal model should be used when the investigator wants to interpret coefficients at the population level and a conditional model should be used when the center is of greater interest. If a population-averaged interpretation is of interest and there is clustering within center, the marginal Cox model should be used. In practice, if correlation within center is even suspected, the marginal Cox model is the best choice. If there is not clustering within center, the traditional Cox model is appropriate. If a center-specific

interpretation is of interest, either the gamma or log normal frailty models can be used. The simulation study results suggest that the gamma and log normal frailty models give very similar results in a variety of settings. The gamma frailty model may be preferred in practice since there is a straight-forward relationship between θ and Kendall's τ. Being able to calculate Kendall's τ easily is useful because it is a statistic that is familiar to a wide range of researchers. However, if these models are going to be used in practice, it would be useful to have some form of model diagnostic to assess the form of the frailty distribution. Some work has been done in this area for parametric models and bivariate models (Shih & Louis, 1995; Oakes, 1989; Duchateau & Janssen, 2008). However, more work needs to be done in this area for semiparametric and multivariate models.

This paper has highlighted the fact that there are different interpretations for marginal and conditional coefficients and that large differences can exist between these two types of coefficients when within center dependence is strong. There is not a good way to perform model selection in this setting as AIC and BIC are not useful, so we offer guidelines that can be used in practice.

Semiparametric Latent Variable Transformation Models for Multiple Outcomes of Mixed Types

Anna C. Snavely, David Harrington, and Yi Li

2.1 Abstract

Multiple outcomes are often collected in applications where the quantity of interest cannot be measured directly, or is difficult or expensive to measure. Latent variable models are commonly adopted in this setting. These models stipulate that the multiple outcomes are conditionally independent measures of the latent factor, possibly capturing various aspects of it. Mixed types of outcomes (e.g. continuous vs discrete) and censored outcomes present statistical challenges, however, as a natural multivariate distribution of mixed data does not exist. In this paper we propose a new class of semiparametric latent variable models that allows for the estimation of the latent factor in the presence of measurable outcomes of mixed types, including censored outcomes. Compared to the existing methods, our proposed model provides the following advantages. First, the model allows the relationship between the measurable outcomes and latent variable to be unspecified, rendering more robust inference. Second, the proposed model can directly estimate the treatment (or other covariate) effect on the unobserved latent variable, greatly enhancing the interpretability of the model. Extensive simulations verify the utility of the methods. We also apply the method to a clinical trial conducted by Dana-Farber Cancer Institute, where the focus was to study the effect of treatment on unobservable dysphagia through collected multiple outcomes, which are of mixed types and one of which is subject to censoring.

2.2 Introduction

Multiple outcomes are often collected in applications where the quantity of interest cannot be measured directly, or is difficult or expensive to measure (Dunson, 2006). Latent variable models are commonly adopted in this setting. These models stipulate that the multiple outcomes are conditionally independent measures

of the latent factor, possibly capturing various aspects of it. Mixed types of outcomes (e.g. continuous vs discrete) are common (Pocock et al., 1987) as are censored outcomes. These varying outcome types present statistical challenges as a natural multivariate distribution of mixed data does not exist. For example, in a head and neck cancer (HNC) trial conducted at Dana-Farber Cancer Institute, the investigators wanted to determine the effect of clinical and treatment factors on dysphagia (or difficulty in swallowing) (Chapuy et al., 2011). However, dysphagia was not directly measurable. Instead, three surrogate outcome measures: duration of feeding tube usage, weight loss after treatment, and diet (liquid, soft, etc) were collected. Among them, the first outcome was subject to censoring, while the other two outcomes are of mixed types; weight loss was measured on a continuous scale, while the diet was measured on an ordinal scale. Limited statistical tools for accommodating such complicated data have greatly hampered proper analyses.

When the measurable outcomes are all continuous, the methods are relatively well developed within the latent variable paradigm (Sammel & Ryan, 1996; Roy & Lin, 2000). Some methods also exist in the context of latent variable modeling when the outcomes are of mixed types. For example, Catalano & Ryan (1992), Fitzmaurice & Laird (1995), Sammel et al. (1997), Regan & Catalano (1999), Moustaki & Knott (2000), and Huber et al. (2004) all consider this setting. However, a limitation of these latent variable models for mixed outcomes is that the relationship between the measurable outcomes and the Gaussian latent variable must be known a priori. Since the latent variable is not observed, there is little guidance for the appropriate relationship. If the relationship is misspecified, using the common likelihood approaches leads to biased estimates for the parameters. Also, these methods do not allow for survival or event time outcomes where censoring is present.

In view of all these limitations, we propose a new class of semiparametric latent variable models that allows for the estimation of the latent factor in the pres-

ence of mixed outcomes as well as censored outcomes. Compared to the existing methods, our proposed model provides the following advantages. First, the model allows the relationship between the measurable outcomes and latent variable to be unspecified, rendering more robust inference. Second, the proposed model can directly estimate the treatment (or other covariate) effect on the unobserved latent variable, greatly enhancing the interpretability of the model.

The remainder of the article is organized as follows. Section 2.3 describes the model and Section 2.4 discusses the estimation and inference procedures. Section 2.5 presents simulation results and applies the methodology to analyze the aforementioned HNC data. We conclude with a discussion in Section 2.6.

2.3 Semiparametric Latent Variable Transformation Models

Suppose there are n subjects, each with p distinct measurable outcomes. For simplicity, we will focus on the setting where there is a single outcome that is subject to censoring, though the extension to accommodate multiple censored outcomes is rather straightforward. Without loss of generality we assume that the first measurable outcome is a continuous event time, denoted by T, which can be censored by a competing censoring variable, denoted by C. We further assume that T and C are independent and that C is independent of the covariates. Let $Y_{i1} = \min(T_i, C_i)$ and $\Delta_i = I(Y_{i1} = T_i)$, where $I(\cdot)$ is the indicator function. Then, for each individual i, we observe vectors of covariates $X_{i1}, \cdots, X_{ip}$ (e.g. age and gender) and Z_i (e.g. treatment), a failure indicator Δ_i, and a vector of measurable outcomes $Y_i = (Y_{i1}, \cdots, Y_{ip})^T$. The elements of Y_i are ordered such that the first p_1 elements are continuous (with the first element being the event time), and the remaining $p_2 = p - p_1$ elements are discrete (including, for example, binary, ordinal, or count outcomes).

In order to facilitate joint modeling, the discrete measurable outcomes are linked to underlying continuous variables as in Muthén (1984) and Dunson (2006). Specifically, let Y_{ij}^u be a continuous variable underlying Y_{ij}. Then, for the discrete outcomes, for $Y_{ij} \in \{1, \cdots, d_j\}$, we have $Y_{ij} = \sum_{l=1}^{d_j} lI(c_j(l-1) < Y_{ij}^u \leq c_j(l))$ where d_j is the number of categories for the jth outcome and $c_j = (c_j(0), \cdots, c_j(d_j))^T$ are unknown thresholds satisfying $-\infty = c_j(0) < \cdots < c_j(d_j) = \infty$. For the measurable outcomes that are already continuous, $Y_{ij} = Y_{ij}^u$.

Given the above notation, we can now relate the continuous or underlying continuous outcomes to the latent variable (e_i) of primary interest through a semiparametric linear transformation model:

$$\begin{aligned} H_1(T_i) &= X_{i1}^T\beta_1 + \alpha_1 e_i + \varepsilon_{i1}, \\ H_2(Y_{i2}^u) &= X_{i2}^T\beta_2 + \alpha_2 e_i + \varepsilon_{i2}, \\ &\vdots \\ H_p(Y_{ip}^u) &= X_{ip}^T\beta_p + \alpha_p e_i + \varepsilon_{ip}. \end{aligned} \tag{2.1}$$

H_1 is an unknown non-decreasing transformation function such that $H(0) = -\infty$ and $H_2, \cdots, H_p$ are unknown non-decreasing transformation functions that satisfy $H_j(-\infty) = -\infty$ and $H_j(\infty) = \infty$ for $j = 2, \cdots, p$. $\beta = (\beta_1^T, \cdots, \beta_p^T)^T$ is a vector of regression coefficients, $\alpha = (\alpha_1, \cdots, \alpha_p)^T$ are factor loadings, e_i is a latent variable for subject i, and $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{ip})^T$ is a vector of independent errors distributed as $N(0, \text{diag}(\sigma_1^2, \cdots, \sigma_p^2))$.

Furthermore, additional structure for the latent variable is assumed:

$$e_i = Z_i^T\gamma + \epsilon_i, \tag{2.2}$$

where Z_i records treatment or other covariates, γ is a vector of unknown regression coefficients, and ϵ_i is the random error distributed as $N(0, \sigma_e^2)$. In most instances, γ is the primary parameter for inference since it relates covariates of interest, such as treatment, to the latent variable (outcome of interest). We assume that Z_i and ϵ_i are

independent and that for identifiability, Z_i and X_{ij} do not contain constant terms, $\sigma_e^2 = 1$, and $\sigma_j^2 = 1$ for $j = 1, \cdots, p$. One of the factor loadings is also constrained to be positive (Dunson, 2003). Though related, our model is different from the ordinary random effect models. Random effects are mainly introduced to describe the unobserved heterogeneity and are usually covariate-independent, whereas the latent variable, e_i, represents specific traits measured by covariates and hence are covariate-dependent.

2.4 Estimation and Inference Procedures

2.4.1 Likelihood and Estimating Equations

For each given $y_j \in \{1, \cdots, d_j\}$ for $j = p_1 + 1, \cdots, p$ (the discrete measurable outcomes), let $\tilde{H}_j(y_j) = H_j(c_j(y_j))$, where c_j is the unknown upper limit of Y_{ij}^u when $Y_{ij} = y_j$. Because both H_j and Y_{ij}^u are unknown, they cannot be identified separately. However, $H_j(c_j(1)), \cdots, H_j(c_j(d_j - 1))$ provide the distribution of the observed outcome Y_{ij} and can be estimated. In other words, for the discrete measurable outcomes, estimation of the transformation means estimation of the unknown transformed thresholds. Also, let $\tilde{H}_j = H_j$ for the continuous measurable outcomes (for ease of notation), $\mathbf{\Theta} = (\beta, \alpha, \gamma)$, and $\tilde{\mathbf{H}} = (\tilde{H}_1, \cdots, \tilde{H}_p)$.

Since the error terms in models (2.1) and (2.2) are assumed to be normally distributed, the vector of transformed continuous outcomes follows a multivariate normal distribution. The likelihood is not simply a multivariate normal density, however, because not all of the outcomes are completely observed. More specifi-

cally, the likelihood can be expressed as:

$$L(\mathbf{\Theta}; \tilde{\mathbf{H}}) \propto |\Sigma_{22}|^{n/2} \prod_{i=1}^{n} \int_{\mathbf{x}^{[2]} \in \mathcal{H}_i^{[2]}} \exp\left[-\frac{1}{2} \left(\begin{pmatrix} \tilde{\mathbf{H}}_i^{[1]} \\ \mathbf{x}^{[2]} \end{pmatrix} - X_i\beta - \alpha\gamma^T Z_i \right)^T \Sigma_{22}^{-1} \left(\begin{pmatrix} \tilde{\mathbf{H}}_i^{[1]} \\ \mathbf{x}^{[2]} \end{pmatrix} - X_i\beta - \alpha\gamma^T Z_i \right) \right] d\mathbf{x}^{[2]}, \quad (2.3)$$

where $X_i = \text{diag}(X_{i1}^T, \cdots, X_{ip}^T)$, $\tilde{\mathbf{H}}_i^{[1]} = (\tilde{H}_1(Y_{i1}^u), \cdots, \tilde{H}_{p_1}(Y_{ip_1}^u))^T$, $\tilde{\mathbf{H}}_i^{[2]} = (\tilde{H}_{p_1+1}(Y_{i,p_1+1}^u), \cdots, \tilde{H}_p(Y_{ip}^u))^T$, and $\mathcal{H}_i^{[2]} = \prod_{j=p_1+1}^{p} [\tilde{H}_j(Y_{ij}), \tilde{H}_j(Y_{ij} + 1)]$. This likelihood arises from the fact that based on models (2.1) and (2.2), $\tilde{\mathbf{H}}_i \equiv (\tilde{\mathbf{H}}_i^{[1]^T}, \tilde{\mathbf{H}}_i^{[2]^T})^T \sim N(X_i\beta + \alpha\gamma^T Z_i, \Sigma_{22})$, where $\Sigma_{22} = \alpha\alpha^T + I_{pxp}$. Here $\tilde{\mathbf{H}}_i^{[1]}$ is completely observed (as long as the event time is not censored), whereas $\tilde{\mathbf{H}}_i^{[2]}$ is only known to fall in $\mathcal{H}_i^{[2]}$. In the case of a censored event time, the event time can be incorporated in $\tilde{\mathbf{H}}_i^{[2]}$. The bounds of integration for the event time ($[\tilde{H}_1(Y_{i1}), \infty]$) can be included in $\mathcal{H}_i^{[2]}$ (since now the time is not completely observed).

The likelihood (specifically the conditional likelihood on $\tilde{\mathbf{H}}$) in equation (2.3) involves the unknown transformation functions, so we need a way to estimate these transformations. Following the usual counting process notation, let $Y_i(t) = I(Y_{i1} \geq t)$ and $N_i(t) = \Delta_i I(Y_{i1} \leq t)$. Then $\tilde{H}_1$ for the event time outcome can be estimated using the following equation (Chen et al., 2002):

$$\sum_{i=1}^{n} \left[dN_i(t) - Y_i(t) d\Lambda\{\tilde{H}_1(t) - X_{i1}^T\beta_1 - \alpha_1 Z_i^T\gamma\} \right] = 0 \quad (t \geq 0), \quad (2.4)$$

where Λ is the cumulative hazard function for the transformed event time (i.e. the cumulative hazard for $N(0, \alpha_1^2 + 1)$). For computational purposes, the following simpler (but asymptotically equivalent) estimating equations can be used:

$$\begin{pmatrix} 1 - \sum_{i=1}^{n} Y_i(t_1)\Lambda\{\tilde{H}_1(t_1) - X_i\beta - \alpha\gamma^T Z_i\} \\ 1 - \sum_{i=1}^{n} Y_i(t_2)\lambda\{\tilde{H}_1(t_2-) - X_i\beta - \alpha\gamma^T Z_i\}\Delta\tilde{H}_1(t_2) \\ \vdots \\ 1 - \sum_{i=1}^{n} Y_i(t_K)\lambda\{\tilde{H}_1(t_K-) - X_i\beta - \alpha\gamma^T Z_i\}\Delta\tilde{H}_1(t_K) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (2.5)$$

where $\Delta\tilde{H}_1(t) = \tilde{H}_1(t) - \tilde{H}_1(t-)$ and K is the number of observed events. The resulting estimate of $\tilde{H}_1(\cdot)$ will be a non-decreasing step function that jumps only at the K observed event times.

$\tilde{H}_j(y_j)$ for $j = 2, \cdots, p$ can be estimated using the following equations:

$$\sum_{i=1}^{n} \left[I(Y_{ij} \leq y_j) - \Phi \left(\frac{\tilde{H}_j(y_j) - (X_{ij}^T \beta_j + \alpha_j Z_i^T \gamma)}{\sqrt{\alpha_j^2 + 1}} \right) \right] = 0 \tag{2.6}$$

where Φ is the standard normal cumulative distribution function.

The estimator $\hat{\tilde{H}}_j(\cdot)$ of $\tilde{H}_j(\cdot)$ is a non-decreasing step function with jumps only at the observed Y_{ij} for the continuous measurable outcomes. For the discrete measurable outcomes, the transformed thresholds are estimated through (2.6). Thus, we have effectively reduced the problem of solving the infinite dimensional system of equations defined by (2.5) and (2.6) to that of solving a finite system of equations.

2.4.2 Estimation Algorithm

We propose a procedure that is similar to a profile likelihood to draw inference. Specifically, given $\mathbf{\Theta}$, the finite dimensional parameters, we use (2.5) and (2.6) to estimate $\tilde{H}_j(\cdot)$ for $j = 1, \cdots, p$ denoted by $\tilde{H}(\mathbf{\Theta})$. We then proceed to estimate $\mathbf{\Theta}$ by maximizing a pseudo-likelihood which is the likelihood function $L(\mathbf{\Theta}, \tilde{\mathbf{H}}(\mathbf{\Theta}))$.

For implementation, we propose the following iterative steps:

Step 1: Choose initial values for β, α, and γ. Denote these estimates by $\hat{\beta}^{(0)}$, $\hat{\alpha}^{(0)}$, and $\hat{\gamma}^{(0)}$. Using an initial estimate of 1 for each of the parameters works well in practice. Picking initial values of 0 for all of the parameters does not work well.

Step 2: Use the estimating equations (2.5) and (2.6) with β, α, and γ set equal to

$\hat{\beta}^{(0)}$, $\hat{\alpha}^{(0)}$, and $\hat{\gamma}^{(0)}$ to obtain initial estimates of the transformation functions, $\hat{\tilde{H}}_j^{(0)}(\cdot)$.

Suppose that we have estimates of β, α, γ, and $\tilde{H}_j(\cdot)$ from the $(m-1)$th iteration; denote these estimates by $\hat{\beta}^{(m-1)}$, $\hat{\alpha}^{(m-1)}$, $\hat{\gamma}^{(m-1)}$, and $\hat{\tilde{H}}_j^{(m-1)}(\cdot)$.

Step 3: Maximize the likelihood (2.3) with respect to β, α, and γ, replacing $\tilde{H}_j(\cdot)$ with $\hat{\tilde{H}}_j^{(m-1)}(\cdot)$, to obtain new estimates: $\hat{\beta}^{(m)}$, $\hat{\alpha}^{(m)}$, and $\hat{\gamma}^{(m)}$. We used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method for maximization in this setting. This is a quasi-Newton method that often performs well for optimization problems (Press et al., 1992).

Step 4: Use the estimating equations (2.5) and (2.6) with β, α, and γ set equal to $\hat{\beta}^{(m)}$, $\hat{\alpha}^{(m)}$, and $\hat{\gamma}^{(m)}$ to obtain new estimates of the transformation functions, $\hat{\tilde{H}}_j^{(m)}(\cdot)$.

Step 5: Repeat *Steps 3* and *4* until predetermined convergence criteria are met.

2.4.3 Bootstrap

To conduct inference on the parameters, standard error estimates are needed. We cannot, however, use the standard error estimates that arise from the likelihood (2.3) as valid estimators of the standard errors for the β, α, and γ parameters. The likelihood estimates do not account for the additional variability that comes from estimating the transformation functions. Instead, we must rely on a resampling procedure to estimate standard errors for the parameters. In the simulations and data analysis we have relied on the traditional nonparametric bootstrap where sampling is done with replacement (Efron & Tibshirani, 1993).

2.5 Numerical Studies

2.5.1 Simulations

We evaluated the performance of the proposed method through simulations. In order to mimic the motivating HNC data, we incorporated three measurable outcomes: an event time outcome, a continuous outcome, and an ordinal outcome with 5 categories. The following model was assumed for the simulations with $H_1 = \log$ and $H_2 = H_3 =$ Identity:

$$\begin{aligned} H_1(T_i) &= X_i\beta_1 + \alpha_1 e_i + \varepsilon_{i1}, \\ H_2(Y_{i2}) &= X_i\beta_2 + \alpha_2 e_i + \varepsilon_{i2}, \\ H_3(Y_{i3}^u) &= X_i\beta_3 + \alpha_3 e_i + \varepsilon_{i3}, \end{aligned} \tag{2.7}$$

and

$$e_i = Z_i\gamma + \epsilon_i. \tag{2.8}$$

Specifically, the underlying continuous variables were generated from the multivariate normal distribution, $N(X_i\beta + \alpha\gamma Z_i, \Sigma_{22})$, where $\Sigma_{22} = \alpha\alpha^T + I_{3\times 3}$. This way the measurable outcomes are correlated, and this correlation is determined by the α parameters. The event time outcome was then created through an anti-log transformation. Censoring was introduced through an exponential random variable, with the parameter for the exponential distribution chosen to given a particular percentage of censoring. The continuous variable did not require further transformation and the ordinal outcome was obtained by using the underlying continuous variable arising from the multivariate normal model and then applying the following thresholds: (-∞, -1, 0, 1, 2, ∞). We assumed that there was a single continuous X covariate common to all three measurable outcomes and a single binary Z covariate to represent treatment or some other binary covariate of interest. Specifically, $X_i \sim N(0, 1)$ and $Z_i \sim \text{Bernoulli}(0.50)$. True parameter values were selected to be $\beta_1 = 0.5$, $\beta_2 = 0.9$, $\beta_3 = 0.75$, $\alpha_1 = 0.5$, $\alpha_2 = 0.9$, $\alpha_3 = 0.75$, $\gamma = 1$.

Six different simulation settings were considered. For each setting, 250 simulations were carried out. Two different sample sizes were explored: n = 100 and n = 200, and three different censoring levels were considered: 0%, 7%, and 17%. The 7% was chosen to mimic the HNC data. Results for n = 100 can be found in Table 2.1 and results for n = 200 are presented in Table 2.2. We discovered through these simulations that there is some numerical instability in the estimation procedure. Convergence of the algorithm is sensitive to the particular data set that you are analyzing. Non-convergence is a definite problem for smaller sample sizes and also for larger amounts of censoring. The percentage of simulations that did not converge for each simulation setting is presented in Table 2.3. The results presented in Table 2.1 and 2.2 and Figure 2.1 and 2.2 are, therefore, conditional on convergence.

Sample size is an important factor in the performance of the proposed method. Figure 2.1 visually compares simulation results for n = 100 and n = 200, both with 7% censoring (since this is the most relevant case for the HNC data). From this figure we can see that the point estimates are not right at the truth for all of the parameters. For example, there does appear to be some underestimation for the parameters associated with the continuous outcome (α_2 and β_2). Despite this fact, none of the parameters are significantly biased since the empirical confidence intervals do not exclude the truth. Also, point estimation can be improved by increasing the sample size. For both sample sizes included in the simulations, the point estimate for the γ parameter is well estimated. This is important because this is the primary parameter for inference since it relates the Z covariate to the latent variable. Inference, however, may be somewhat unreliable for a sample as small as 100. For example, numerical instability of the estimation procedure is more of a problem with a small sample size (21.7% of simulations failed to converge) and the 95% coverage probabilities based on the bootstrap standard errors tend to deviate somewhat from the nominal level. In particular, even though the point estimate for γ seems reasonable for n = 100, the bootstrap standard error is somewhat over-

Table 2.1: Simulation results for n = 100

	β_1	β_2	β_3	α_1	α_2	α_3	γ
0% Censoring							
Mean	0.476	0.713	0.752	0.450	0.576	0.746	1.134
Bias	-0.024	-0.187	0.002	-0.050	-0.324	-0.004	0.134
Empirical SE	0.128	0.117	0.156	0.220	0.183	0.234	0.349
Bootstrap SE	0.125	0.139	0.162	0.200	0.203	0.203	0.363
95% CI Coverage	0.935	0.970	0.959	0.858	0.935	0.911	0.976
7% Censoring							
Mean	0.519	0.666	0.682	0.622	0.525	0.574	0.979
Bias	0.019	-0.234	-0.068	0.122	-0.375	-0.176	-0.021
Empirical SE	0.187	0.150	0.162	0.381	0.234	0.256	0.316
Bootstrap SE	0.152	0.151	0.164	0.289	0.210	0.226	0.411
95% CI Coverage	0.938	0.928	0.959	0.851	0.877	0.892	0.995
17% Censoring							
Mean	0.845	0.570	0.603	1.889	0.504	0.332	0.522
Bias	0.345	-0.330	-0.147	1.389	-0.396	-0.418	-0.478
Empirical SE	0.529	0.179	0.137	1.502	0.174	0.148	0.428
Bootstrap SE	0.274	0.173	0.146	0.712	0.178	0.173	0.446
95% CI Coverage	0.810	0.946	0.958	0.744	0.958	0.976	0.952

Table 2.2: Simulation results for n = 200

	β_1	β_2	β_3	α_1	α_2	α_3	γ
0% Censoring							
Mean	0.492	0.770	0.773	0.492	0.683	0.759	1.083
Bias	-0.008	-0.130	0.023	-0.008	-0.217	0.009	0.083
Empirical SE	0.090	0.092	0.128	0.140	0.116	0.171	0.238
Bootstrap SE	0.088	0.107	0.123	0.152	0.151	0.164	0.250
95% CI Coverage	0.942	0.971	0.947	0.938	0.988	0.951	0.959
7% Censoring							
Mean	0.524	0.730	0.694	0.650	0.609	0.595	1.007
Bias	0.024	-0.170	-0.056	0.150	-0.291	-0.155	0.007
Empirical SE	0.120	0.112	0.111	0.324	0.181	0.187	0.294
Bootstrap SE	0.106	0.119	0.118	0.231	0.171	0.176	0.276
95% CI Coverage	0.915	0.969	0.964	0.906	0.924	0.942	0.920
17% Censoring							
Mean	0.987	0.599	0.599	2.180	0.534	0.350	0.478
Bias	0.487	-0.301	-0.151	1.680	-0.366	-0.400	-0.522
Empirical SE	0.536	0.143	0.092	1.499	0.167	0.140	0.301
Bootstrap SE	0.260	0.139	0.100	0.708	0.156	0.139	0.323
95% CI Coverage	0.947	0.965	0.941	0.971	0.935	0.900	0.953

Table 2.3: Percentage of simulations that failed to converge

	0% Censoring	7% Censoring	17% Censoring
n = 100	32.4	21.7	31.7
n = 200	2.8	10.4	31.7

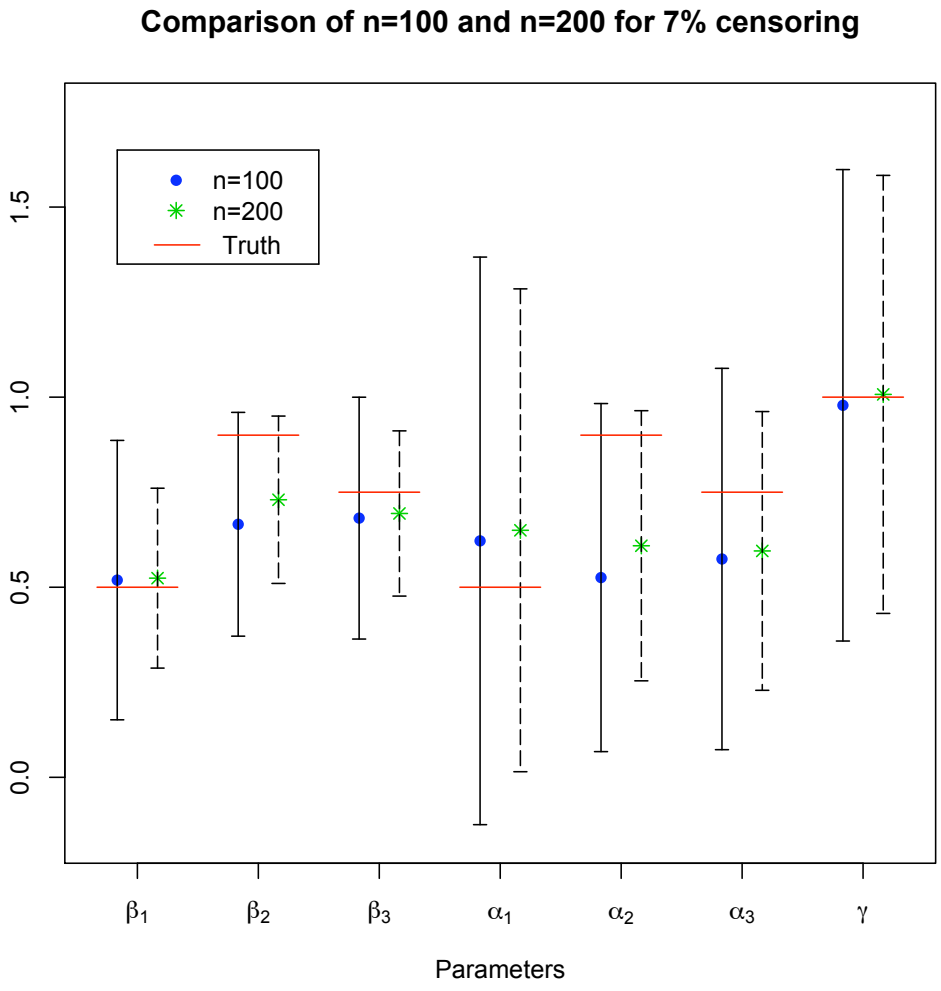

Figure 2.1: Plots of means and empirical 95% confidence intervals from simulations with n = 100 and n = 200 with 7% censoring.

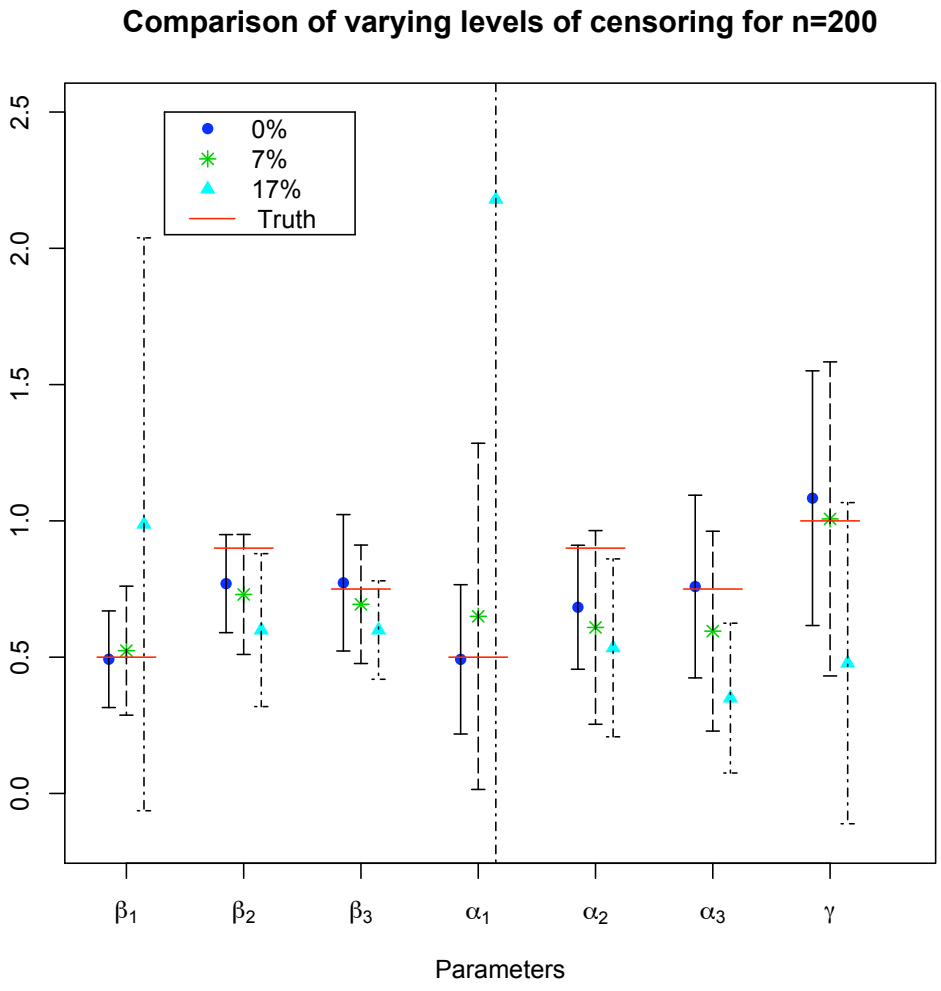

Figure 2.2: Plots of means and empirical 95% confidence intervals from simulations with n = 200 and 0%, 7%, and 17% censoring.

estimated, leading to a coverage probability that is too large. On the other hand, inference for n = 200 with 7% censoring appears to be reliable as demonstrated by 95% coverage probabilities that are close to the nominal level. The larger sample size of 200 also has the added advantage of better numerical stability, with only 10.4% of simulations failing to converge.

Censoring has a strong impact of the performance of the proposed methodology. Figure 2.2 considers simulation results for 0%, 7%, and 17% censoring for the larger sample size of 200. The larger sample size was chosen due to the better numerical stability. It is clear from the figure that when the censoring reaches the moderate level of 17%, performance of the proposed method suffers. The standard errors of the parameters associated with the event time (α_1 and β_1) are huge and there are parameters that are significantly biased based on the empirical confidence intervals. Even the γ parameter is not well estimated with a larger amount of censoring. But, when there is no censoring, the performance of the proposed method is good. Only 2.8% of the simulations did not converge and the coverage probabilities are close to the nominal level.

Simulation results, therefore, suggest that the proposed methodology can be quite useful when there is a larger sample size with a small amount of censoring. However, care should be taken when the percentage of censoring is high or when the sample size is small.

2.5.2 DFCI Head and Neck Data

We applied the proposed method to a study of head and neck cancer (HNC) patients carried out at Dana-Farber Cancer Institute (Chapuy et al., 2011). Patients were identified for the study through a retrospective chart review and were eligible if they were diagnosed between 1998 and 2008 with an advanced-stage squamous cell carcinoma of the oropharynx, hypopharynx, larynx, or unknown primary and

were treated with chemoradiotherapy (chemoRT) and neck dissection. Twenty-four months of follow-up after chemoRT without recurrence was also required for inclusion. The goal of the study is to determine clinical and treatment factors associated with dysphagia in this group of HNC patients. Dysphagia, or difficulty swallowing, is a major concern for HNC patients after treatment because it can have a negative impact on a patient's quality of life. However, there is not one definitive way to measure or define the condition objectively and physicians often do not want to define the condition subjectively based on patients' perceptions of swallowing. Caudell et al. (2009) recognized the usefulness of utilizing multiple objective measures to describe dysphagia. They created a single composite endpoint using several objective measures they felt would capture all patients suffering from dysphagia (they believed some patients might be missed using only a single outcome measure). Chapuy et al. (2011) also used multiple outcomes to create a score used in the analysis. While the idea of using multiple measurable outcomes in this way may be an improvement over using a single outcome, it is not ideal because the composite endpoint or score is defined somewhat arbitrarily.

The investigators at Dana-Farber collected information on several measurable outcomes that are often used to describe dysphagia: time from end of chemoRT to removal of the gastrostomy tube, weight loss after chemoRT, and diet (liquid, soft, etc). Using this information they want to identify factors associated with dysphagia, not factors associated with a single outcome related to dysphagia. The analysis of this data set is challenging because the outcome of interest, dysphagia, is not observable and we have multiple measurable outcomes of mixed types available that are all attempting to measure dysphagia that we would like to combine in a meaningful way to evaluate treatment and other factors. Using the proposed methodology for the analysis will allow us to combine the multiple measurable outcomes through the latent variable structure and then determine factors associated with the latent variable (dysphagia) as desired by the investigators.

Eighty-eight patients were eligible for the study. Two patients were excluded from the analysis because they never had a gastrostomy tube and thus represent a different patient population. Sixty-six patients were then available with complete outcome information. We were able to impute weight information using a later weight measurement (measurement taken at some point after our baseline of 1 month post chemoRT) for 9 patients, giving us 75 patients available for analysis. For each patient i, let Y_{i1} be the observed time from end of chemoRT to removal of the gastrostomy tube in days (note this outcome is potentially censored), Y_{i2} be weight loss after chemoRT in kg, and Y_{i3} be diet (regular, soft, pureed, liquid, no food; ordinal 1-5). α_3 will be constrained to be greater than 0 for identifiability. Using the proposed methodology, e_i characterizes the level of dysphagia for patient i with a larger e_i indicating worse dysphagia and Z_i is treatment or some other clinical factor potentially associated with dysphagia. In this way, γ is the parameter of primary interest for inference.

When analyzing the HNC data set, we are hindered by the small sample size ($n = 75$). We are unable to fit complex models with many covariates, so instead we will focus on two models that are small, but clinically interesting. Model 1 will include T-stage (ordinal) as the Z covariate and sex as the X covariate and Model 2 will include treatment (induction vs. concurrent chemoRT) as the Z covariate and sex as the X covariate. T-stage is clinically relevant because T-stage has been shown to be associated with adverse swallowing outcomes previously (Machtay et al., 2008; Nguyen et al., 2009; Chapuy et al., 2011). Treatment is of interest to determine if patients treated with induction chemotherapy followed by chemoRT have worse dysphagia as compared to patients treated with primary concurrent chemoRT. Sex is included as the X covariate in both models because it is not a variable of primary interest, but we might want to control for it in the analysis.

Results for Model 1 and Model 2 can be found in Table 2.4. For both models, none of the β parameters are significant as evidenced by 95% confidence intervals

Table 2.4: Head and neck data analysis results

	Estimate	SE	95% CI
Model 1 (T-stage)			
β_1	0.154	0.413	(-0.655, 0.964)
β_2	-0.582	0.301	(-1.171, 0.008)
β_3	-0.220	0.466	(-1.135, 0.694)
α_1	0.577	0.288	(0.012, 1.141)
α_2	0.017	0.127	(-0.233, 0.266)
α_3	1.518	0.614	(0.315, 2.721)
γ	-0.027	0.356	(-0.724, 0.670)
Model 2 (Treatment)			
β_1	0.167	0.423	(-0.663, 0.997)
β_2	-0.581	0.367	(-1.301, 0.138)
β_3	-0.156	0.670	(-1.469, 1.158)
α_1	0.528	0.146	(0.242, 0.815)
α_2	0.019	0.094	(-0.166, 0.204)
α_3	1.341	0.304	(0.746, 1.936)
γ	0.231	0.401	(-0.554, 1.016)

*For both models sex is the X covariate

that cover 0. This suggests that sex is not associated with any of the transformed outcomes included in the model. Also, for both models α_1 and α_3 are significant, but α_2 is not. The α parameters are factor loadings, so these findings indicate that time on the gastrostomy tube and diet are significantly associated with the latent variable (dysphagia). In fact, worse dysphagia is associated with a longer time on the feeding tube and a more modified diet. Weight loss after chemoRT, however, does not appear to be significantly related to dysphagia. This is not a surprising result as clinically we know that weight loss may not be a great measure of dysphagia. On the one hand, it makes sense that if a patient is having difficulty swallowing, then that patient is likely to eat less and lose more weight. However, when the feeding tube is being used, adequate nutrition can be obtained through the tube without the need to swallow. This would suggest that even if swallowing is difficult for a patient, this may not be seen through measuring the weight of that patient. Both Model 1 and Model 2 indicate that weight loss may not be a useful measure to capture dysphagia.

In Model 1, T-stage is included as the Z covariate. The γ parameter associated with T-stage in this model is not significant. This suggests that a higher T-stage (increased size of the primary tumor) is not associated with worse dysphagia. Similarly, from Model 2, there is no evidence of an association between treatment and dysphagia, meaning that there is not evidence that induction chemotherapy is associated with worse dysphagia. While neither of the Z covariates considered were found to be significant, it is important to keep in mind the limitations of the proposed methodology when interpreting these results. The sample size for the HNC data is 75 and there is 6.7% censoring. From the simulation results, we know that inference may not be totally reliable in this setting. In particular, the standard error estimate for the γ parameter may be too large. Therefore, in a larger data set, we might find an association between one or both of the Z covariates and dysphagia. In this way, a larger sample size would be useful in order to be more certain

about results obtained using the proposed methodology.

2.6 Discussion

The proposed methodology has a lot of appeal because it uses a semiparametric approach that does not require pre-specifying a link between the measurable outcomes and the latent variable of interest. The methodology also has the advantage of allowing for multiple outcomes of mixed types, including censored outcomes, to be incorporated into the latent variable framework and being able to estimate the treatment (or other covariate) effect on the unobserved latent variable. Simulations suggest that the proposed method has a lot of utility when the sample size is large and the censoring proportion is small. More specifically, the method performs well for a sample size of 200 with 7% censoring or less. However, there are some limitations to this approach that must be acknowledged. In particular, when the censoring percentage reaches 17%, simulations indicate inference may not reliable. Also, sample size is a concern. Specifically, a sample size of 100 may be too small to completely trust inference made using the proposed method. As noted previously, there is evidence of some numerical instability in the estimation algorithm when the sample size is small. A larger sample size is likely required because of the large number of parameters that must be estimated, either in the transformation functions or through the likelihood. In the end, a small sample likely does not contain enough information to reliably estimate all of the pieces. Despite these limitations, a sample size of 200 or more is not unreasonable in many potential applications.

Another potential limitation of the proposed method is the use of a threshold model for discrete measurable outcomes. The threshold model is practical because then we have all continuous outcomes that can be jointly modeled using the multivariate normal. However, this means that nominal measurable outcomes cannot be meaningfully incorporated in this approach. Also, in order to use the

threshold approach, it must be plausible that there is some underlying continuous quantity that gives rise to the ordinal categories that are observed. In the HNC case, the ordinal outcome is diet and it is plausible that there is some underlying biological quantity that determines the change from one food type to another. However, it is possible the threshold approach would not make sense in a different application. Choosing which covariates should be X covariates and which should be Z covariates is also an issue. In general, if you want to be able to look at the association between a covariate and the latent variable, then that covariate should be included in Z. However, there is not a good statistical approach to make this decision.

This methodology was motivated by the question of what treatment and clinical factors are associated with dysphagia in HNC patients. We are limited in our ability to address this clinical question, however, by having only 75 patients available for analysis. We were able to fit two simple models to the HNC data, but it is not clear how meaningful these results would be in the clinical literature. Despite the limitations of the proposed method in small samples, the semiparametric approach could have great utility in informing a parametric model of the same form as models (2.1) and (2.2), but with the link functions pre-specified. In particular, we can use the estimated transformations that arise out of the proposed methodology to decide on appropriate link functions for a parametric model. Figure 2.3 presents the estimated transformations obtained from Model 1. Based on this plot, it would seem reasonable to use either a log or square root transformation for the time on the gastrostomy tube. Using the identity link (i.e. assuming normality) for weight loss also seems appropriate. As discussed previously, the semiparametric method proposed in this paper is a useful approach for analysis when the sample size is large. In addition, we suggest that the proposed method has a very practical use, even in a smaller sample, in informing a parametric version of the method.

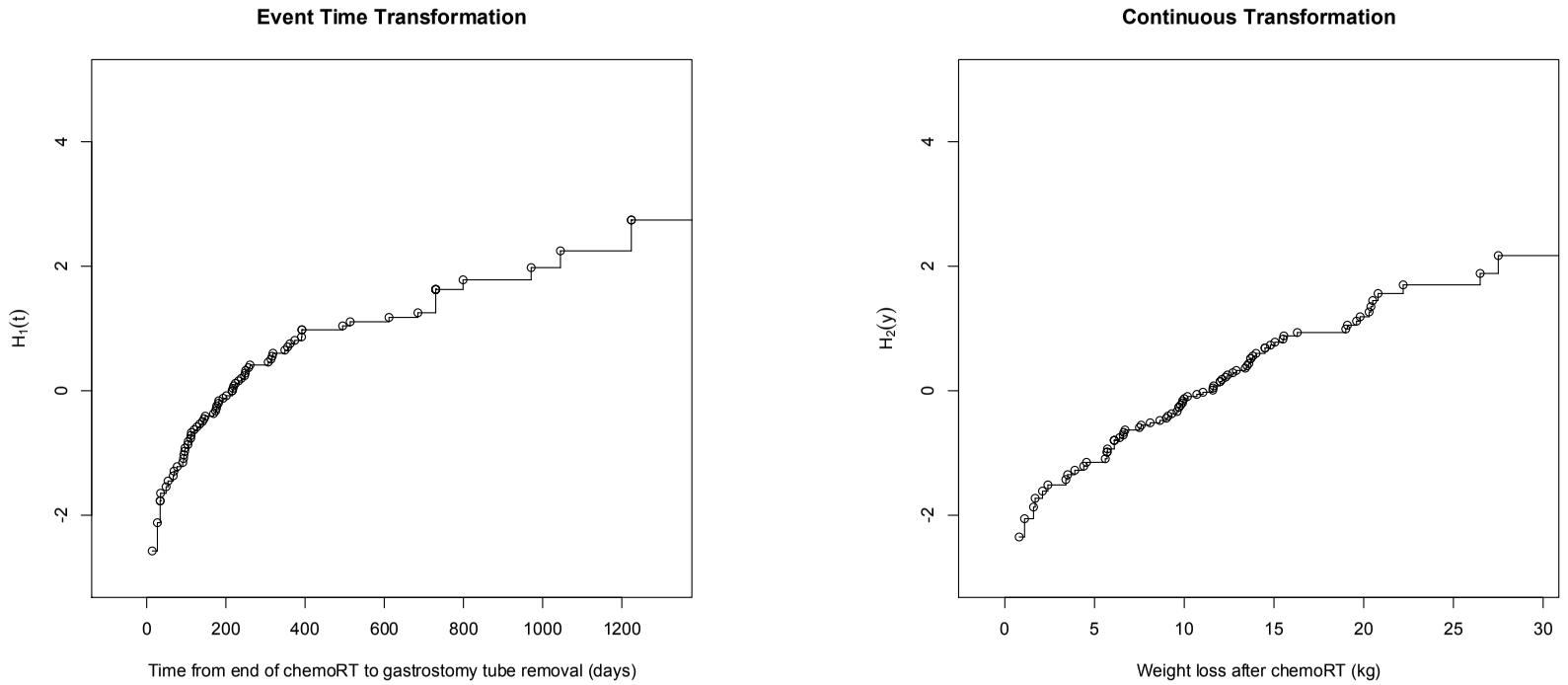

Figure 2.3: Estimated transformations for Model 1 (T-stage)

Exploring Dysphagia in Head and Neck Cancer Patients: A Latent Variable Transformation Model Approach

Anna C. Snavely, David Harrington, and Yi Li

3.1 Abstract

Dysphagia, or difficulty swallowing, is a common issue for head and neck cancer patients post treatment. Quality of life is negatively impacted by swallowing difficulty, so it is important to understand clinical and treatment factors that might be related to the condition. The challenge here is that dysphagia, our outcome of interest, cannot be measured directly. Instead, there are multiple measurable outcomes of mixed types available that are all attempting to capture some aspect of dysphagia. In this paper we propose a latent variable normal transformation model where the measurable outcomes are assumed to be governed by an unobserved (latent) variable, which in turn may depend on covariates such as treatment. This approach has the advantage of incorporating measurable outcomes that are both discrete and continuous. In particular, the proposed method extends traditional methods by including measurable outcomes that are subject to censoring. Through the structure of the model we are able to study the effect of covariates on the latent variable, which in our case represents dysphagia, greatly enhancing the interpretability of the model. The methodology is applied to a study of head and neck cancer patients from Dana-Farber Cancer Institute.

3.2 Introduction

Squamous cell carcinoma of the head and neck represents about 5% of newly diagnosed cancers in adults in the United States. These patients tend to present with locally advanced disease and are treated aggressively with some combination of surgery, chemotherapy, and radiotherapy (Posner et al., 2007). Intensive chemoradiation (chemoRT) regimens have been found to be effective in the management of head and neck cancer (HNC) in terms of improving both progression free and overall survival (Pignon et al., 2000). With aggressive chemoRT treatment, how-

ever, comes side effects such as dysphagia, or difficulty swallowing, that have a negative impact on a patient's quality of life (Goguen et al., 2006; Nguyen et al., 2009). Having a neck dissection (surgical procedure to remove lymph nodes in the neck) following chemoRT may have an additional impact on swallowing function as seen in Machtay et al. (2008) and Lango et al. (2010). With more patients surviving after aggressive treatment, understanding what factors may be related to dysphagia becomes important as quality of life becomes a critical consideration for both patients and physicians.

Investigators at Dana-Farber Cancer Institute (DFCI) carried out a study with the goal of exploring treatment and clinical factors that may be associated with dysphagia (Chapuy et al., 2011). This goal is challenging, however, because there is not one definitive way to measure or define dysphagia objectively. Physicians often do not want to define the condition subjectively based on patients' perceptions of swallowing because patient reported outcomes have not been found to be well correlated with more objective measures such as videofluoroscopy findings (Caudell et al., 2009). In order to best capture dysphagia, the use of multiple objective measures has been suggested (Caudell et al., 2009; Chapuy et al., 2011). This has resulted in analysis based on a single composite endpoint or score based on objective measures. This approach may be an improvement over using a single outcome, but it is not ideal because the composite endpoint or score is defined somewhat arbitrarily. Instead, we propose a latent variable normal transformation model to handle this analysis. In the proposed model, measurable outcomes of mixed types (including event times) are assumed to be governed by an unobserved (latent) variable, which in turn may depend on covariates such as treatment. This approach has the advantage of incorporating measurable outcomes that are both discrete and continuous. In particular, the proposed method extends traditional methods by including measurable outcomes that are subject to censoring. Also, through the structure of the model we are able to study the effect of covariates on

the latent variable, greatly enhancing the interpretability of the model. In the HNC setting, the measurable outcomes are the objective measures available for capturing dysphagia and the latent variable represents dysphagia itself, our outcome of interest. This approach, therefore, will allow us to combine the multiple measurable outcomes in a meaningful way through the latent variable structure and then determine factors associated with the latent variable (dysphagia) as desired by the investigators.

The particulars of the DFCI study will be described in Section 3.3 and the model will be specified and discussed in Section 3.4. Simulation results will be presented in Section 3.5, followed by data analysis results in Section 3.6. We will wrap-up with a discussion in Section 3.7.

3.3 DFCI Head and Neck Study

Investigators at Dana-Farber conducted a retrospective study to learn about dysphagia in HNC patients treated with chemoRT and neck dissection (Chapuy et al., 2011). Both electronic and paper medical records were reviewed to identify eligible patients. Patients were included in the study if they were diagnosed with an advanced-stage squamous cell carcinoma of the oropharynx, hypopharynx, larynx, or unknown primary and were treated at DFCI between January 1998 and June 2008 with chemoRT and neck dissection. All patients had at least twenty-four months of follow-up after chemoRT without recurrence and had a primary site complete response to chemoRT at the time of neck dissection. The 88 patients found to be eligible for the study, therefore, represent a group of patients who had advanced disease and responded well to aggressive treatment (no recurrence for at least two years). Quality of life is an important consideration for this subset of patients, so learning more about dysphagia in this group is particularly relevant.

Since there is no direct objective measure for dysphagia, the DFCI investigators collected information on several measurable outcomes that can be used as surrogate measures for dysphagia: time from end of chemoRT to removal of the gastrostomy tube, weight loss after chemoRT, and diet (regular, soft, pureed, liquid, no food). Time from end of chemoRT to gastrostomy tube (feeding tube directly into the stomach) removal is an event time subject to censoring. The censoring in this case is administrative (i.e. feeding tube still in place at the end of follow-up). We would expect a longer time on the feeding tube to be indicative of worse dysphagia. Weight loss after chemoRT is continuous. We might expect that a greater weight loss would be associated with worse dysphagia. However, this may not be true if the patient is getting adequate nutrition through the feeding tube. In this way, it is not clear that weight loss is a great surrogate measure for dysphagia. Diet is an ordinal variable that characterizes how modified a patient's diet is, with a more modified diet expected to be related to worse dysphagia. Both weight loss and diet are measured at baseline, which is 1 month after the end of chemoRT. As an important note, because the feeding tube is a gastrostomy tube, it is possible for a patient to still have a feeding tube in place and be able to eat food by mouth. In other words, having a feeding tube in place does not automatically restrict a patient to the "no food" diet category.

Of the 88 patients that were eligible for the study, 75 will be used for the analysis. Two patients were excluded because they never had a gastrostomy tube placed and thus represent a different patient population. After the 2 exclusions, there were 66 patients with complete outcome information. In order to increase the sample size and make use of as much data as possible, we imputed weight information using a later weight measurement (measurement taken at some point after our baseline of 1 month post chemoRT) for 9 patients. This gives us 75 patients available for analysis. This is the same set of patients that is analyzed in Chapter 2.

3.4 Latent Variable Transformation Model

3.4.1 Background and Model Specification

Latent variable models were introduced in the field of psychology with the first methods being attributed to Spearman (1904). Spearman's methodology lead to the development of factor analysis and other latent variable approaches that have been used extensively in both psychology and education. Use of latent variable methods in the biomedical setting has increased over time, particularly as latent variable methodology has been further developed (see, for example, Sammel & Ryan (1996); Roy & Lin (2000); Sammel et al. (1997); Moustaki & Knott (2000); Dunson (2006)).

For the analysis of the DFCI head and neck data, we propose a latent variable model that builds on the basic structure of Sammel & Ryan (1996). The model suggested by Sammel & Ryan (1996) only allows for continuous measurable outcomes, however, so we extend their approach by including discrete outcomes by linking the discrete outcomes to underlying continuous outcomes as originally considered in Muthén (1984). The underlying continuous variables can then be incorporated in the multivariate normal structure. We also add additional flexibility by considering transformed outcomes and allowing for censored measurable outcomes. These extensions will allow us to incorporate events times, continuous outcomes, and ordinal outcomes as measurable outcomes in the latent variable framework as required by the HNC data. Figure 3.1 presents a diagram of the basic model structure. In short, measurable outcomes of mixed types are assumed to be governed by a latent variable (e), which in turn may depend on covariates (Z) such as treatment. Additional covariates, X, may also be related to the measurable outcomes. The actual modeling is done on all continuous variables and transformations, H, are allowed. The transformation functions, however, must be

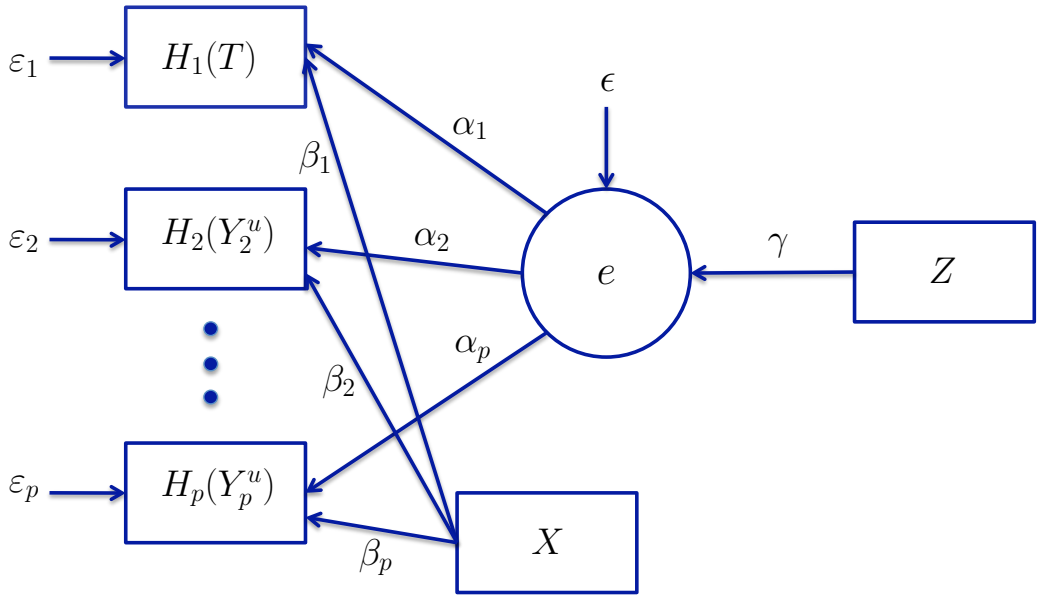

Figure 3.1: Diagram of basic model structure for the latent variable transformation approach.

pre-specified.

Specifically, suppose there are n subjects with p distinct measurable outcomes. For simplicity, we will consider the setting where there is a single measurable outcome that is subject to censoring. The extension to accommodate multiple censored outcomes, however, is rather straightforward. Without loss of generality we will assume that the event time, T, is the first measurable outcome. C is then the corresponding censoring variable. We assume that T and C are independent and also that C is independent of the covariates. For individual i, let $Y_{i1} = \min(T_i, C_i)$ and $\Delta_i = I(Y_{i1} = T_i)$, where $I(\cdot)$ is the indicator function. Then, for each individual, we observe vectors of covariates $X_{i1}, \cdots, X_{ip}$ (e.g. age and gender) and Z_i (e.g. treatment), a failure indicator Δ_i, and a vector of measurable outcomes $Y_i = (Y_{i1}, \cdots, Y_{ip})^T$. The elements of Y_i are ordered such that the first p_1 elements are continuous (with the first element being the event time), and the remaining

$p_2 = p - p_1$ elements are discrete (binary, ordinal, or count outcomes are possible).

In order to facilitate joint modeling, the discrete measurable outcomes are linked to underlying continuous variables as in Muthén (1984) and Dunson (2006). Let Y_{ij}^u be a continuous variable underlying Y_{ij}. Then, for the discrete outcomes, for $Y_{ij} \in \{1, \cdots, d_j\}$, we have $Y_{ij} = \sum_{l=1}^{d_j} lI(c_j(l-1) < Y_{ij}^u \leq c_j(l))$ where d_j is the number of categories for the jth outcome and $c_j = (c_j(0), \cdots, c_j(d_j))^T$ are unknown thresholds satisfying $-\infty = c_j(0) < \cdots < c_j(d_j) = \infty$. For the measurable outcomes that are already continuous, Y_{ij} is simply equal to Y_{ij}^u.

The continuous or underlying continuous outcomes can now be related to the latent variable (e_i) of primary interest through the following model:

$$\begin{aligned} H_1(T_i) &= X_{i1}^T \beta_1 + \alpha_1 e_i + \varepsilon_{i1}, \\ H_2(Y_{i2}^u) &= X_{i2}^T \beta_2 + \alpha_2 e_i + \varepsilon_{i2}, \\ &\vdots \\ H_p(Y_{ip}^u) &= X_{ip}^T \beta_p + \alpha_p e_i + \varepsilon_{ip}. \end{aligned} \tag{3.1}$$

In model (3.1), $\beta = (\beta_1^T, \cdots, \beta_p^T)^T$ is a vector of regression coefficients, $\alpha = (\alpha_1, \cdots, \alpha_p)^T$ are factor loadings, e_i is a latent variable for subject i, and $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{ip})^T$ is a vector of independent errors distributed as $N(0, \text{diag}(\sigma_1^2, \cdots, \sigma_p^2))$. The Hs are pre-specified monotone transformation functions. A typical choice for the event time would a log transformation. The identity link may be an appropriate choice for many continuous measurable outcomes (i.e. assume normality). For the discrete measurable outcomes, a transformation does not need to be specified since we do not observe the underlying continuous variable. Transformed thresholds ($H_j(c_j(1)), \cdots, H_j(c_j(d_j - 1))$) can be estimated through the likelihood, where the assumption is that the underlying continuous variable has been transformed to be normally distributed.

The latent variable is assumed to have additional structure as follows:

$$e_i = Z_i^T \gamma + \epsilon_i, \tag{3.2}$$

where Z_i records covariates of interest such as treatment, γ is a vector of unknown regression coefficients, and ϵ_i is the random error distributed as $N(0, \sigma_e^2)$. In most instances, γ is the primary parameter for inference since it relates important covariates to the latent variable (outcome of interest). We assume that Z_i and ϵ_i are independent and that for identifiability, Z_i and X_{ij} do not contain constant terms, $\sigma_e^2 = 1$, and $\sigma_j^2 = 1$ for $j = p_1 + 1, \cdots, p$ (discrete measurable outcomes). One of the factor loadings is also constrained to be positive (Dunson, 2003).

3.4.2 Likelihood Specification and Parameter Estimation

Since the error terms in models (3.1) and (3.2) are assumed to be normally distributed, the vector of transformed continuous outcomes follows a multivariate normal distribution. The likelihood, however, is not simply a multivariate normal density because not all of the outcomes are completely observed. For each given $y_j \in \{1, \cdots, d_j\}$ for $j = p_1 + 1, \cdots, p$ (the discrete measurable outcomes), let $\tilde{H}_j(y_j) = H_j(c_j(y_j))$, where c_j is the unknown upper limit of Y_{ij}^u when $Y_{ij} = y_j$. For the continuous measurable outcomes, let $\tilde{H}_j = H_j$. Then if $\mathbf{\Theta} = (\beta, \alpha, \gamma)$ and $\tilde{\mathbf{H}} = (\tilde{H}_1, \cdots, \tilde{H}_p)$, the likelihood can be expressed as:

$$L(\mathbf{\Theta}; \tilde{\mathbf{H}}) \propto |\Sigma_{22}|^{n/2} \prod_{i=1}^{n} \int_{\mathbf{x}^{[2]} \in \mathcal{H}_i^{[2]}} \exp\left[-\frac{1}{2} \left(\begin{pmatrix} \tilde{\mathbf{H}}_i^{[1]} \\ \mathbf{x}^{[2]} \end{pmatrix} - X_i\beta - \alpha\gamma^T Z_i \right)^T \Sigma_{22}^{-1} \left(\begin{pmatrix} \tilde{\mathbf{H}}_i^{[1]} \\ \mathbf{x}^{[2]} \end{pmatrix} - X_i\beta - \alpha\gamma^T Z_i \right) \right] d\mathbf{x}^{[2]}, \tag{3.3}$$

where $X_i = \text{diag}(X_{i1}^T, \cdots, X_{ip}^T)$, $\tilde{\mathbf{H}}_i^{[1]} = (\tilde{H}_1(Y_{i1}^u), \cdots, \tilde{H}_{p_1}(Y_{ip_1}^u))^T$, $\tilde{\mathbf{H}}_i^{[2]} = (\tilde{H}_{p_1+1}(Y_{i,p_1+1}^u), \cdots, \tilde{H}_p(Y_{ip}^u))^T$, and $\mathcal{H}_i^{[2]} = \prod_{j=p_1+1}^{p} [\tilde{H}_j(Y_{ij}), \tilde{H}_j(Y_{ij} + 1)]$.

The likelihood (3.3) arises from the fact that based on models (3.1) and (3.2), $\tilde{\mathbf{H}}_i \equiv (\tilde{\mathbf{H}}_i^{[1]^T}, \tilde{\mathbf{H}}_i^{[2]^T})^T \sim N(X_i\beta + \alpha\gamma^T Z_i, \Sigma_{22})$, where $\Sigma_{22} = \alpha\alpha^T + \Psi$,

$\Psi = \text{diag}(\sigma_1^2, \cdots, \sigma_p^2))$, and $\sigma_j^2 = 1$ for $j = p_1 + 1, \cdots, p$ (discrete measurable outcomes). The integrals in the likelihood are required because not all of the measurable outcomes are completely observed. In particular, $\tilde{\mathbf{H}}_i^{[1]}$ is completely observed (as long as the event time is not censored), however $\tilde{\mathbf{H}}_i^{[2]}$ is only known to fall in $\mathcal{H}_i^{[2]}$. In the case of a censored event time, the event time can also be incorporated in $\tilde{\mathbf{H}}_i^{[2]}$ and the bounds of integration for the event time ($[\tilde{H}_1(Y_{i1}), \infty]$) can be included in $\mathcal{H}_i^{[2]}$ (since now the time is not completely observed).

The model parameters most relevant for inference are β, α, and γ. However, the transformed thresholds associated with the discrete measurable outcomes and the variance parameters associated with the continuous measurable outcomes also need to be estimated. All of these parameters can be estimated through maximizing the likelihood (3.3).

Many different maximization routines could be considered, but we have opted to use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. BFGS is a quasi-Newton method that often performs well for complex optimization problems (Press et al., 1992). Whatever method is used, however, must support a constrained maximization. For each discrete measurable outcome, the transformed thresholds must be constrained to be ordered, and one of the factor loadings (i.e. one of the α parameters) must be constrained to be positive. Because this is a constrained optimization problem, inference based on likelihood theory could be incorrect. However, simulation results suggest that the model based standard errors arising from the likelihood are reliable and, therefore, inference based on these standard errors is reasonable. In other words, the constraints do not appear to cause boundary issues in this setting and traditional likelihood theory can be used for inference.

3.4.3 Model Checking

The transformation functions, H, must be pre-specified in order to use this parametric latent variable transformation model. Choosing an appropriate transformation can be a challenge, however, so it would be useful to have guidance in selecting a transformation and a means of diagnosing a misspecified transformation.

Chapter 2 presented a semiparametric latent variable transformation model. In this approach, the transformations are estimated using the data instead of being pre-specified. Estimated transformations (in the form of step functions) can be obtained through the estimation procedure. These step functions can then be used to inform a parametric model. For example, the shape of an estimated transformation from the semiparametric approach might suggest using a log link in a parametric model.

After a parametric model has been fit, residuals can be used to consider model fit. Residuals can be obtained separately for each of the measurable outcomes. In the continuous case the residuals will be: $H_j(y_{ij}) - X_{ij}^T\hat{\beta}_j + \hat{\alpha}_j Z_i^T\hat{\gamma}$. In other words, the residuals are simply the observed transformed outcome minus the predicted transformed outcome from the model. Similar to the linear regression setting, these residuals should be normally distributed and have mean 0. Also, we would not expect to see a pattern in the residuals in a plot of the residuals vs. the fitted values if the model fits well. A misspecified transformation should be noticeable in the residual plots. Figures 3.2 and 3.3 show residual plots for a correctly specified model. Both the density and Q-Q plots suggest that there are no major departures from normality and the Residuals vs. Fitted plot shows no recognizable pattern.

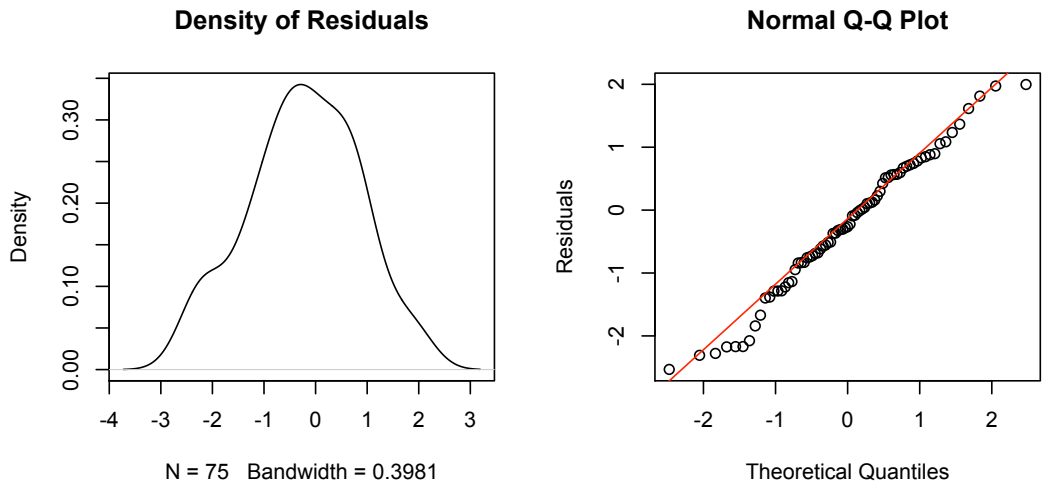

Figure 3.2: Plot looking at the normality of the residuals for the event time for the correct model (log link).

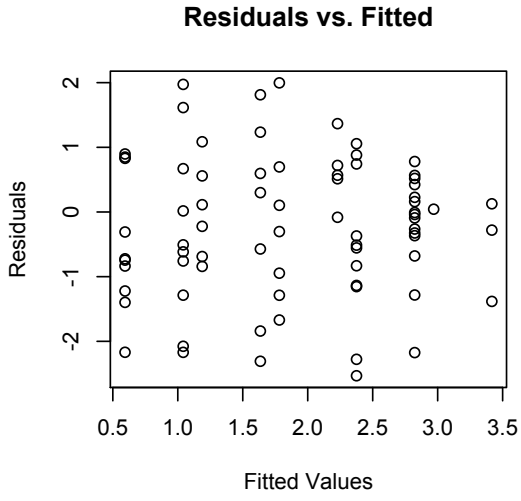

Figure 3.3: Plot of residuals vs. fitted values for the event time for the correct model (log link).

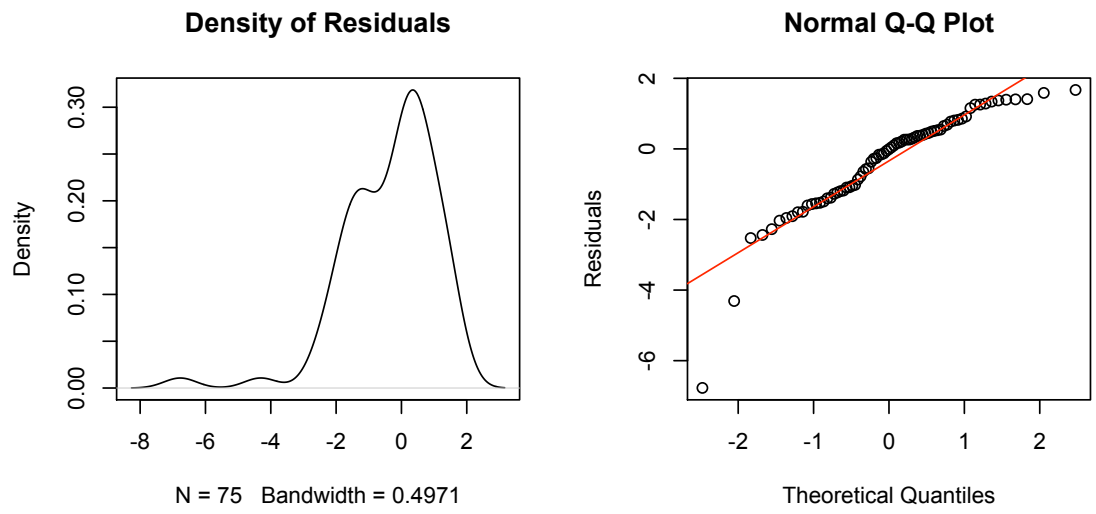

Figure 3.4: Plot looking at the normality of the residuals for the event time when the transformation is misspecified (square root instead of log link).

Figures 3.4 and 3.5 show residual plots for a model where the event time transformation is misspecified. The true transformation is a square root, but the model assumes a log link. Clear departures from normality are seen by the long left tail in the density plot and the curved shape of the residuals in the Q-Q plot. The Residuals vs. Fitted plot shows a few residuals that are very small. These residual plots can be very useful is diagnosing an incorrect transformation and in general, such plots are useful in assessing model fit.

For the discrete measurable outcomes, the underlying continuous variable can be predicted from the model by $X_{ij}^T\hat{\beta}_j + \hat{\alpha}_j Z_i^T\hat{\gamma}$. The estimated transformed thresholds can then be applied to the underlying continuous variables to get a predicted category for each individual, i. Model fit could then be assessed by seeing how well the predicted categories and observed categories match up in a cross-classified table. Association could be assessed using a Fisher's Exact test, but there is limited power for this kind of test, particularly in a small sample.

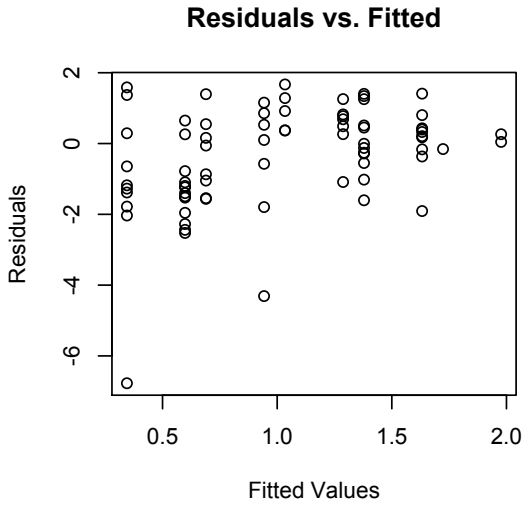

Figure 3.5: Plot of residuals vs. fitted values for the event time when the transformation is misspecified (square root instead of log link).

3.5 Simulation Studies

We used simulations to evaluate the performance of the latent variable transformation model. Since this methodology was proposed to analyze the DFCI head and neck data, the simulations were designed to mimic the HNC data structure. Three measurable outcomes were, therefore, considered in the simulations: an event time outcome, a continuous outcome, and an ordinal outcome with 5 categories. The following model was fit to the simulated data:

$$\begin{aligned} \log(T_i) &= X_i\beta_1 + \alpha_1 e_i + \varepsilon_{i1}, \\ Y_{i2} &= X_i\beta_2 + \alpha_2 e_i + \varepsilon_{i2}, \\ Y_{i3}^u &= X_i\beta_3 + \alpha_3 e_i + \varepsilon_{i3}, \end{aligned} \tag{3.4}$$

and

$$e_i = Z_i\gamma + \epsilon_i. \tag{3.5}$$

In the simulations considered we assumed that there was a single continuous X covariate common to all three measurable outcomes and a single binary Z

covariate to represent treatment or some other binary covariate of interest. Specifically, $X_i \sim N(0,1)$, $Z_i \sim \text{Bernoulli}(0.50)$, and the true parameter values were selected to be $\beta_1 = 0.5$, $\beta_2 = 0.9$, $\beta_3 = 0.75$, $\alpha_1 = 0.5$, $\alpha_2 = 0.9$, $\alpha_3 = 0.75$, $\gamma = 1$. The underlying continuous variables were then generated from the multivariate normal, $N(X_i\beta + \alpha\gamma Z_i, \Sigma_{22})$, where $\Sigma_{22} = \alpha\alpha^T + \Psi$. $\Psi = \text{diag}(\sigma_1^2, \sigma_2^2, 1)$ and we assume that $\sigma_1 = 1$ and $\sigma_2 = 1$, or in other words, $\Sigma_{22} = \alpha\alpha^T + I_{3\times 3}$. Generating the data in this way ensures that the measurable outcomes are correlated through the α parameters. For the correctly specified model, the event time outcome was then created through an anti-log transformation with censoring introduced through an exponential random variable. The parameter for the exponential distribution was chosen to give a particular percentage of censoring. The continuous outcome did not require further transformation for the correctly specified model. Finally, the ordinal outcome was obtained by using the underlying continuous variable arising from the multivariate normal model and then applying the following thresholds: (-∞, -1, 0, 1, 2, ∞).

3.5.1 Correctly Specified Model Results

Eight simulation settings were considered to explore performance of the latent variable transformation approach. For each setting, 250 simulations were carried out. Two different sample sizes were explored: n = 100 and n = 200, and four different censoring levels were considered: 0%, 7%, 17%, and 50%. The 7% censoring was chosen to match the HNC data and the 50% censoring was chosen to illustrate a higher censoring setting. Primary parameter results (β, α, and γ) for n = 100 can be found in Table 3.1. Corresponding results for n = 200 are presented in Table 3.2. Secondary parameter results (transformed thresholds and standard deviation parameters) for both sample sizes are included in Table 3.3.

Overall, simulation results suggest that the performance of the latent vari-

Table 3.1: Simulation results for n = 100 (primary parameters)

	β_1	β_2	β_3	α_1	α_2	α_3	γ
0% Censoring							
Mean	0.504	0.900	0.825	0.497	0.902	0.852	1.042
Bias	0.004	<0.001	0.075	-0.003	0.002	0.102	0.042
Empirical SE	0.121	0.132	0.522	0.110	0.161	0.897	0.251
Model SE	0.112	0.135	0.198	0.109	0.158	0.254	0.246
95% CI Coverage	0.952	0.944	0.964	0.956	0.960	0.968	0.952
7% Censoring							
Mean	0.497	0.912	0.795	0.480	0.876	0.788	1.065
Bias	-0.003	0.012	0.045	-0.020	-0.024	0.038	0.065
Empirical SE	0.114	0.137	0.173	0.107	0.169	0.223	0.275
Model SE	0.113	0.133	0.183	0.110	0.157	0.240	0.249
95% CI Coverage	0.952	0.956	0.972	0.968	0.940	0.968	0.920
17% Censoring							
Mean	0.493	0.880	0.786	0.494	0.882	0.826	1.030
Bias	-0.007	-0.020	0.036	-0.006	-0.018	0.076	0.030
Empirical SE	0.118	0.147	0.198	0.121	0.141	0.269	0.240
Model SE	0.116	0.134	0.203	0.116	0.159	0.283	0.244
95% CI Coverage	0.948	0.932	0.952	0.936	0.976	0.956	0.940
50% Censoring							
Mean	0.493	0.881	0.779	0.497	0.885	0.811	1.028
Bias	-0.007	-0.019	0.029	-0.003	-0.015	0.061	0.028
Empirical SE	0.138	0.148	0.178	0.147	0.148	0.239	0.239
Model SE	0.131	0.134	0.202	0.136	0.168	0.285	0.250
95% CI Coverage	0.951	0.930	0.963	0.942	0.971	0.988	0.963

Table 3.2: Simulation results for n = 200 (primary parameters)

	β_1	β_2	β_3	α_1	α_2	α_3	γ
0% Censoring							
Mean	0.496	0.906	0.787	0.494	0.888	0.774	1.030
Bias	-0.004	0.006	0.037	-0.006	-0.012	0.024	0.030
Empirical SE	0.074	0.098	0.125	0.078	0.110	0.173	0.180
Model SE	0.079	0.095	0.123	0.077	0.111	0.160	0.169
95% CI Coverage	0.964	0.948	0.944	0.952	0.960	0.944	0.920
7% Censoring							
Mean	0.506	0.903	0.781	0.500	0.900	0.780	1.020
Bias	0.006	0.003	0.031	$<$0.001	$<$0.001	0.030	0.020
Empirical SE	0.082	0.101	0.132	0.073	0.118	0.162	0.168
Model SE	0.080	0.096	0.123	0.078	0.110	0.159	0.167
95% CI Coverage	0.944	0.948	0.932	0.960	0.928	0.944	0.956
17% Censoring							
Mean	0.492	0.906	0.772	0.487	0.889	0.768	1.020
Bias	-0.008	0.006	0.022	-0.013	-0.011	0.018	0.020
Empirical SE	0.084	0.099	0.117	0.078	0.105	0.164	0.172
Model SE	0.082	0.095	0.121	0.081	0.113	0.159	0.170
95% CI Coverage	0.940	0.936	0.968	0.956	0.964	0.936	0.940
50% Censoring							
Mean	0.489	0.906	0.770	0.486	0.894	0.764	1.017
Bias	-0.011	0.006	0.020	-0.014	-0.006	0.014	0.017
Empirical SE	0.100	0.099	0.117	0.093	0.112	0.167	0.183
Model SE	0.093	0.095	0.122	0.095	0.118	0.163	0.174
95% CI Coverage	0.928	0.936	0.968	0.948	0.964	0.952	0.928

Table 3.3: Simulation results - estimates of secondary parameters

	$H_3(c_3(1))$	$H_3(c_3(2))$	$H_3(c_3(3))$	$H_3(c_3(4))$	σ_1	σ_2
n=100						
0% Censoring	-1.064	0.035	1.125	2.234	0.886	0.926
7% Censoring	-1.029	0.028	1.075	2.132	0.905	0.910
17% Censoring	-1.061	0.014	1.079	2.151	0.926	0.855
50% Censoring	-1.061	0.007	1.064	2.127	0.965	0.919
n=200						
0% Censoring	-1.031	0.009	1.038	2.063	0.989	0.995
7% Censoring	-1.013	0.011	1.036	2.082	0.992	0.978
17% Censoring	-1.029	-0.002	1.033	2.053	0.992	0.982
50% Censoring	-1.029	-0.005	1.028	2.048	0.983	0.976

*True thresholds are: -1, 0, 1, 2. True standard deviations are 1.

able transformation model is quite good in a variety of settings. In particular, the bias for the β, α, and γ parameters is consistently small and the coverage probabilities all tend to be close to the nominal level. A higher censoring percentage does not create estimation difficulties, but rather simply increases standard errors a bit. The model based standard errors tend to be quite close to the empirical standard errors, suggesting that inference based on these model based standard errors should be reliable. The noticeable exception to this is for the β_3 and α_3 parameters for n = 100. In this case, the model based standard errors are small relative to the empirical standard errors. The large empirical standard errors are driven by 2 simulations with extreme values. When the sample size is smaller, there are occasional situations ($\sim$1%) when the maximum likelihood estimation fails (seen by standard error estimates that are listed as NA) or where results are not reliable (seen by exceptionally large parameter estimates). These situations, however, should be easily recognizable by an analyst and should be treated as a model that cannot be reliably fit.

Table 3.4: Simulation results for n = 75 with 7% censoring to mimic HNC data

	β_1	β_2	β_3	α_1	α_2	α_3	γ
Mean	0.512	0.913	0.836	0.503	0.879	0.834	1.050
Bias	0.012	0.013	0.086	0.003	-0.021	0.084	0.050
Empirical SE	0.126	0.151	0.285	0.132	0.189	0.479	0.316
Model SE	0.131	0.155	0.235	0.130	0.182	0.318	0.299
95% CI Coverage	0.959	0.951	0.955	0.963	0.959	0.939	0.943

	$H_3(c_3(1))$	$H_3(c_3(2))$	$H_3(c_3(3))$	$H_3(c_3(4))$	σ_1	σ_2
Mean	-1.084	0.012	1.107	2.254	0.931	0.926

*True thresholds are: -1, 0, 1, 2. True standard deviations are 1.

The benefits of an increased sample size are illustrated in Table 3.3. The transformed thresholds tend to be somewhat better estimated in a larger sample and the standard deviation estimates are definitely improved in a larger sample. A larger sample is, therefore, preferable. However, results in Table 3.1 suggest that even with a smaller sample size, inference of the primary parameters is quite reasonable.

To further evaluate performance of the method for the DFCI head and neck data specifically, we considered a simulation with n = 75 and censoring of 7%. Results for this setting can be found in Table 3.4. Results in this case are very similar to the setting when n = 100. The biggest change is that there are a few more simulations where the maximum likelihood estimation fails, but this still only represents a little over 2% of simulations. This occasional instability likely arises from the fact that we are trying to estimate 13 parameters with only 75 subjects. Overall, however, the latent variable transformation model seems like a reasonable analysis approach for the HNC data as long as maximum likelihood estimation does not fail for the particular model being fit.

3.5.2 Misspecified Model Results

Because the transformations must be pre-specified in this approach, we may be particularly interested in the impact on parameter estimates when the transformations are misspecified. The event time outcome usually will require some transformation. A log link would be a logical choice, but may not always be the appropriate link function. In Tables 3.5, 3.6, and 3.7 we present simulation results for the setting where a log link is fit to the data, but a square root link should have been used. In other words, the event time data were generated through a square transformation instead of an anti-log transformation.

For the larger sample size of 200, we see that the parameters associated with the event time (β_1, α_1, and σ_1) are biased. However, this is to be expected since these parameters would now have a different interpretation. The other parameters, however, are still well estimated. In particular, γ still has a fairly small bias and good coverage probability. This would suggest that misspecifying the event time link in this way should not have a major impact on the inference for γ, which is of primary interest for answering the clinical question. When the sample size is decreased to 100, the estimation of γ, as well as β_3 and α_3 (the parameters associated with the ordinal measurable outcome), is not quite as good. There is a bit larger absolute bias for these parameters than with the larger sample size, though the coverage probabilities are still quite good. Also, the smaller sample size with misspecification of the event time link leads to a bit more instability in the maximum likelihood procedure. The percentage of simulations that cannot be estimated through maximum likelihood, however, is still under 7%.

We also considered the scenario when both the event time and continuous links were misspecified. The event time link was misspecified in the same way as before. The continuous measurable outcome was generated using an anti-log transformation, but the identity link was fit to the data. Results for this setting are

Table 3.5: Simulation results for n = 100 (primary parameters); event time link misspecified

	β_1	β_2	β_3	α_1	α_2	α_3	γ
0% Censoring							
Mean	0.185	0.901	0.892	-0.253	0.871	0.948	1.175
Bias	-0.315	0.001	0.142	-0.753	-0.029	0.198	0.175
Empirical SE	0.235	0.133	1.096	0.226	0.229	1.796	0.811
Model SE	0.231	0.135	0.219	0.217	0.205	0.324	0.348
95% CI Coverage	0.942	0.942	0.959	0.950	0.913	0.954	0.959
7% Censoring							
Mean	0.170	0.907	0.939	-0.199	0.856	1.014	1.138
Bias	-0.330	0.007	0.189	-0.699	-0.044	0.264	0.138
Empirical SE	0.236	0.136	1.206	0.230	0.235	1.968	0.382
Model SE	0.239	0.132	0.235	0.226	0.199	0.334	0.355
95% CI Coverage	0.962	0.954	0.950	0.941	0.908	0.941	0.941
17% Censoring							
Mean	0.149	0.883	0.924	-0.181	0.859	1.047	1.101
Bias	-0.351	-0.017	0.174	-0.681	-0.041	0.297	0.101
Empirical SE	0.254	0.148	1.242	0.237	0.199	2.169	0.307
Model SE	0.252	0.134	0.244	0.243	0.199	0.351	0.315
95% CI Coverage	0.944	0.927	0.962	0.953	0.940	0.970	0.944
50% Censoring							
Mean	0.175	0.879	0.881	0.034	0.894	0.994	1.036
Bias	-0.25	-0.021	0.131	-0.466	-0.006	0.244	0.036
Empirical SE	0.338	0.148	0.820	0.322	0.194	1.487	0.271
Model SE	0.330	0.134	0.248	0.329	0.209	0.374	0.287
95% CI Coverage	0.951	0.930	0.959	0.951	0.959	0.971	0.959

Table 3.6: Simulation results for n = 200 (primary parameters); event time link misspecified

	β_1	β_2	β_3	α_1	α_2	α_3	γ
0% Censoring							
Mean	0.161	0.905	0.842	-0.249	0.840	0.846	1.102
Bias	-0.339	0.005	0.092	-0.749	-0.060	0.096	0.102
Empirical SE	0.153	0.098	0.734	0.158	0.187	1.028	0.299
Model SE	0.163	0.095	0.143	0.156	0.144	0.215	0.226
95% CI Coverage	0.972	0.955	0.967	0.931	0.907	0.963	0.939
7% Censoring							
Mean	0.148	0.903	0.788	-0.213	0.878	0.788	1.072
Bias	-0.352	0.003	0.038	-0.713	-0.022	0.038	0.072
Empirical SE	0.161	0.102	0.148	0.160	0.164	0.219	0.215
Model SE	0.172	0.096	0.140	0.164	0.145	0.207	0.208
95% CI Coverage	0.972	0.948	0.952	0.944	0.907	0.964	0.944
17% Censoring							
Mean	0.169	0.906	0.778	-0.214	0.864	0.770	1.076
Bias	-0.331	0.006	0.028	-0.714	-0.036	0.020	0.076
Empirical SE	0.182	0.100	0.126	0.163	0.153	0.187	0.230
Model SE	0.178	0.094	0.132	0.172	0.145	0.196	0.210
95% CI Coverage	0.935	0.935	0.955	0.967	0.943	0.963	0.955
50% Censoring							
Mean	0.198	0.906	0.776	-0.023	0.900	0.770	1.022
Bias	-0.302	0.006	0.026	-0.523	$<$0.001	0.020	0.022
Empirical SE	0.237	0.099	0.127	0.218	0.154	0.211	0.209
Model SE	0.230	0.095	0.134	0.225	0.149	0.198	0.198
95% CI Coverage	0.944	0.936	0.968	0.980	0.932	0.952	0.956

Table 3.7: Simulation results - estimates of secondary parameters; event time link misspecified

	$H_3(c_3(1))$	$H_3(c_3(2))$	$H_3(c_3(3))$	$H_3(c_3(4))$	σ_1	σ_2
n=100						
0% Censoring	-1.073	0.116	1.287	2.474	2.063	0.825
7% Censoring	-1.165	0.085	1.323	2.568	2.196	0.774
17% Censoring	-1.174	0.092	1.335	2.557	2.267	0.762
50% Censoring	-1.171	0.048	1.248	2.459	2.754	0.849
n=200						
0% Censoring	-1.056	0.058	1.150	2.226	2.115	0.910
7% Censoring	-0.992	0.044	1.080	2.132	2.201	0.959
17% Censoring	-1.008	0.027	1.068	2.092	2.311	0.945
50% Censoring	-1.033	-0.001	1.040	2.068	2.706	0.925

*True thresholds are: -1, 0, 1, 2. True standard deviations are 1.

not shown, but it is important to note that this amount of misspecification leads to a fairly unstable maximum likelihood procedure (maximum likelihood estimation failed for as many as 30% of simulations) and leads to substantial bias in all of the parameters. The use of the semiparametric procedure to inform a parametric model and the use of residuals should help to avoid this scenario, however.

3.6 DFCI Head and Neck Data Analysis

We have 75 patients available for analysis from the DFCI head and neck study. Because we are dealing with a rather small sample size, we are limited in the complexity of the models that can be fit to the data. Also, from the simulation studies we know that when there is a small sample size maximum likelihood estimation may fail for some models. Despite these limitations, we were able to fit a number of models to the HNC data.

All models utilized the same three measurable outcomes, but differed in the covariates included and the link function assumed for the event time. Specifically, the measurable outcomes are defined as follows for patient i: Y_{i1} is the observed time from end of chemoRT to removal of the gastrostomy tube in days (note this outcome is potentially censored), Y_{i2} is weight loss after chemoRT in kg, and Y_{i3} is diet (regular, soft, pureed, liquid, no food; ordinal 1-5). For identifiability, α_3 is constrained to be greater than 0. Using the latent variable transformation structure, e_i characterizes the level of dysphagia for patient i with a larger e_i indicating worse dysphagia. Z_i is treatment or some other clinical factor of interest that is potentially associated with dysphagia. In this way, γ is the parameter of primary interest for inference.

We found that using a log link for the time on the feeding tube and an identity link for weight loss after chemoRT provided the best performance. After considering a number of different models with varying covariates, the model that best fit the data did not include any X covariates, but included both treatment and T-stage as Z covariates. Both treatment (Z_{i1}) and T-stage (Z_{i2}) were of particular interest to the DFCI investigators. Treatment here is an indicator of having induction chemotherapy in addition to chemoRT. Treatment is of interest to determine if patients treated with the more aggressive induction chemotherapy followed by chemoRT have worse dysphagia as compared to patients treated with primary concurrent chemoRT. T-stage is clinically relevant because it has been shown to be associated with adverse swallowing outcomes in the literature (Machtay et al., 2008; Nguyen et al., 2009; Chapuy et al., 2011).

Results for this model can be found in Table 3.8. In particular all of the factor loadings are significant, suggesting that all three of our surrogate measures are contributing information about dysphagia. Specifically, a longer time on the feeding tube is associated with worse dysphagia, more weight loss after chemoRT is associated with worse dysphagia, and a more modified diet is associated with

Table 3.8: Final head and neck results including treatment (Z_{i1}) and T-stage (Z_{i2})

	Estimate	SE	95% CI
α_1	2.311	0.207	(1.904, 2.717)
α_2	4.713	0.520	(3.695, 5.732)
α_3	1.884	0.354	(1.190, 2.578)
γ_1	1.099	0.240	(0.629, 1.569)
γ_2	0.629	0.096	(0.441, 0.817)

worse dysphagia. γ_1 captures the relationship between treatment and dysphagia. Results suggest that after accounting for T-stage, having induction chemotherapy is associated with worse dysphagia. In other words, the more aggressive treatment does seem to be related to more swallowing difficulty. Similarly, γ_2 looks at the relationship between T-stage and dysphagia. After accounting for treatment, a higher T-stage is associated with worse dysphagia. When the same model is fit assuming a square root link for time on the feeding tube, all of the parameters remain significant so we would draw the same conclusions. In addition, the γ parameter estimates are very close for both models, suggesting that the results are robust to the particular link in this setting.

In selecting a final model, we utilized residual plots to consider model fit. The semiparametric analysis from Chapter 2 produced transformation estimates that suggested a log or square root link might be reasonable for the time on the feeding tube and that the identity link would reasonable for weight loss. For the time on the feeding tube, we fit both models assuming a log link and assuming a square root link. Residual plots suggested that the log link was a better fit, although as noted above this choice did not influence the γ parameter estimates much. Residual plots were also used to assess model fit in general. For example, consider the residuals for time on the feeding tube from the model only including treatment (Figure 3.6). There is a clear departure from normality, suggesting that the model with treatment only does not fit very well. The residuals for weight loss

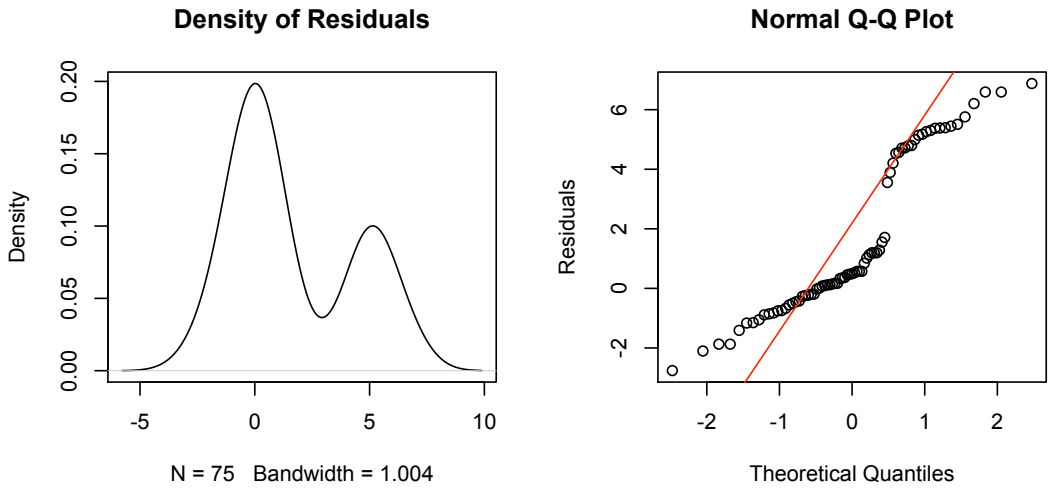

Figure 3.6: Plot looking at the normality of the residuals for time on the feeding tube with only treatment included in the model.

also demonstrated departure from normality, although the departure was not quite as extreme. Residuals for the model including both treatment and T-stage can be seen in Figures 3.7 and 3.8. From Figure 3.7 we see that there is huge improvement in the model fit with the addition of T-stage. The density and Q-Q plots for both of the continuous measurable outcomes suggest a pretty good fit. In particular, we do not see a pattern in these residuals that would suggest that the link function is incorrectly specified. Figure 3.8, however, does show a clear pattern in the Residuals vs. Fitted plots. This suggests that there is likely a covariate missing from the analysis. However, due to the small sample size and limited covariates available for analysis, this is still the best fitting model.

The Swallowing Performance Scale (SPS) score as determined from a video swallow study is another measure sometimes used to capture dysphagia. Ideally, we would have the SPS score at baseline for all of the patients in the study so that SPS score could have been used as an additional measurable outcome. However, this information was not available due to the retrospective nature of the study. We do have SPS score information for 48 patients (though the timing of the measure is

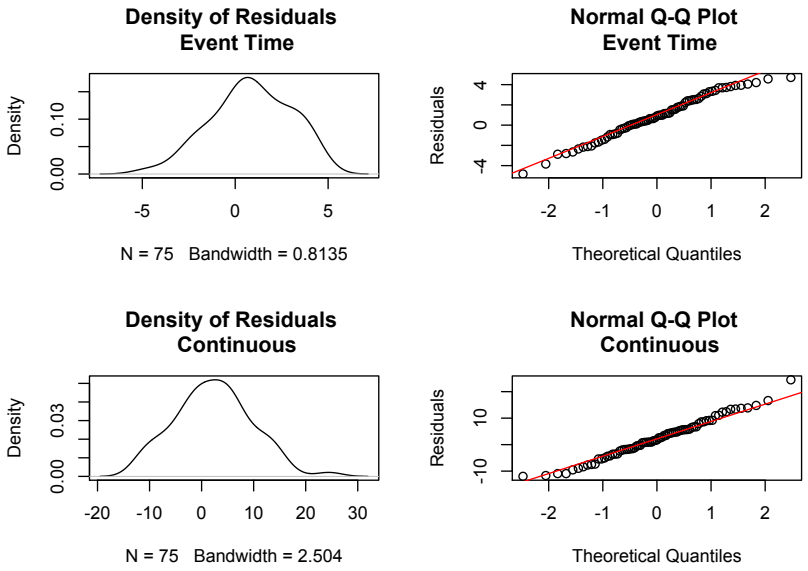

Figure 3.7: Plots looking at the normality of the residuals for model including both treatment and T-stage.

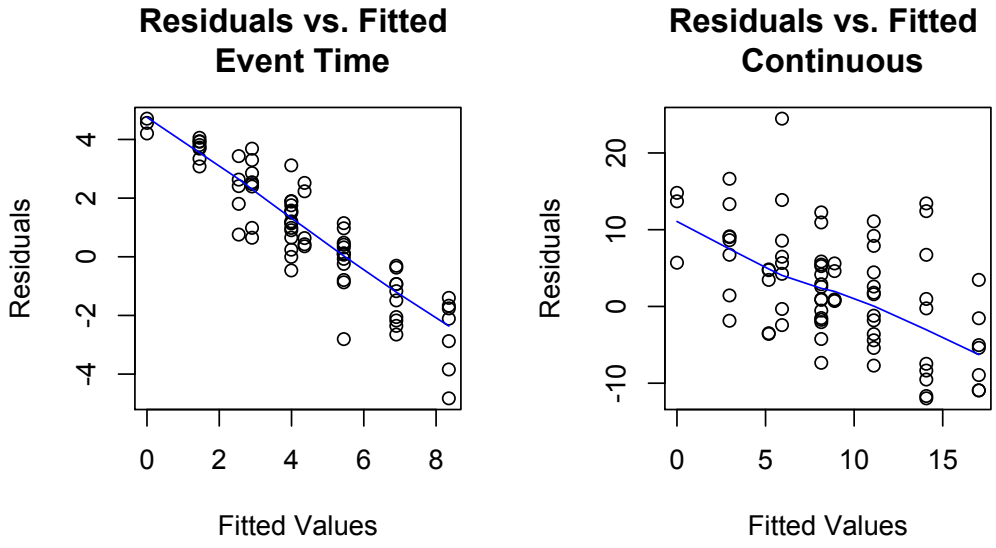

Figure 3.8: Plots of residuals vs. fitted values for model including both treatment and T-stage.

all over the map). As an additional check of our model, we could look at the correlation between the predicted latent variables (in this case $Z_i^T\hat{\gamma}$) and the SPS score for the subset of patients with an SPS score. If our latent variable is really capturing dysphagia we would expect there to be a positive correlation, as a higher SPS score means worse dysphagia. We found the correlation to be 0.40. This suggests that our latent variable is at least doing a reasonable job of capturing dysphagia.

3.7 Discussion

The latent variable transformation model that we have proposed for the HNC analysis is a useful approach for the DFCI data in particular, but also as a general latent variable method. Specifically, the model we have proposed has the advantage of being able to incorporate event times subject to censoring, continuous outcomes, and discrete outcomes (binary, ordinal, or count) as measurable outcomes in a latent variable framework. In this way, we have covered all potential outcome types except for nominal outcomes. While you must pre-specify link functions in this approach, we have suggested the use of the semiparametric methodology proposed in Chapter 2 to inform a parametric model and have demonstrated the use of residuals to diagnose incorrect transformation functions. The performance of the latent variable transformation model is quite good, even when you have a fairly small sample size, unless you have substantial model misspecification.

In exploring dysphagia, specifically, we are limited by the small sample size and retrospective design of the DFCI study. The small sample size is a general concern because of the number of parameters that must be estimated using this latent variable transformation approach. Not only do the α, β, and γ parameters have to be estimated, but so do the transformed thresholds and standard deviation parameters. When the models can be fit using maximum likelihood, the performance is pretty good, even with a sample as small as 75 as in the DFCI data. However, we

are limited by the sample size in the number of covariates that can be included. Also, there are instances when models of interest simply cannot be fit because the maximum likelihood procedure fails.

In exploring the DFCI data, we did run into models that simply could not be fit and/or covariates that could not be included because of too few people in each category. For example, DFCI investigators were interested in whether type of radiation is related to dysphagia. However, there were just not enough patients in each of the different radiation categories to be able to consider this variable in the model. Also, the model including sex as the X covariate and treatment as the Z covariate is an example of a model that could not be fit due to a failed maximum likelihood procedure. Because the study was retrospective, we also did not have the advantage of the treatment assignment being randomized and were missing covariates such as smoking status and alcohol use that may be particularly relevant to head and neck data.

Despite these limitations, we were able to use a small data set to learn something about dysphagia through the latent variable approach. Receiving induction chemotherapy in addition to chemoRT appears to be associated with worse dysphagia as does a higher T-stage. Also, we have proposed a solid methodological approach that can be useful in other settings and could be used to further investigate dysphagia once more data is collected.

References

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine*, 7, 1121–1137.

Anello, C., O'Neill, R. T., & Dubey, S. (2005). Multicentre trials: a us regulatory perspective. *Statistical Methods in Medical Research*, 14(3), 303 – 318.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24, 1713–1723.

Cai, J. & Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, 82(1), 151–164.

Catalano, P. J. & Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87(419), 651–658.

Caudell, J. J., Schaner, P. E., Meredith, R. F., Locher, J. L., Nabell, L. M., Carroll, W. R., Magnuson, J. S., Spencer, S. A., & Bonner, J. A. (2009). Factors associated with long-term dysphagia after definitive radiotherapy for locally advanced head-and-neck cancer. *International Journal of Radiation Oncology*Biology*Physics*, 73(2), 410 – 415.

Chapuy, C. I., Annino, D. J., Snavely, A., Li, Y., Tishler, R. B., Norris, C. M., Haddad, R. I., & Goguen, L. A. (2011). Swallowing function following postchemoradiotherapy neck dissection. *Otolaryngology – Head and Neck Surgery*, 145(3), 428–434.

Chen, K., Jin, Z., & Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3), 659–668.

Commenges, D. & Andersen, P. K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis*, 1, 145–156.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 74, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2), 269–276.

Duchateau, L. & Janssen, P. (2008). *The Frailty Model*. Springer: New York.

Dunson, D. B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association*, 98(463), pp. 555–563.

Dunson, D. B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostat*, 7(4), 551–568.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Fine, J. P., Glidden, D. V., & Lee, K. E. (2003). A simple estimator for a shared frailty regression model. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(1), 317–329.

Fitzmaurice, G. M. & Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, 90(431), 845–852.

Fleiss, J. L. (1986). Analysis of data from multiclinic trials. *Controlled Clinical Trials*, 7(4), 267 – 275.

Glidden, D. V. & Vittinghoff, E. (2004). Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*, 23(3), 369–388.

Goguen, L. A., Posner, M. R., Norris, C. M., Tishler, R. B., Wirth, L. J., Annino, D. J., Gagne, A., Sullivan, C. A., Sammartino, D. E., & Haddad, R. I. (2006). Dysphagia after sequential chemoradiation therapy for advanced head and neck cancer. *Otolaryngology – Head and Neck Surgery*, 134(6), 916–922.

Grambsch, P. M. & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3), 515–526.

Gray, R. J. (1994). A bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*, 50(1), 244–253.

Gray, R. J. (1995). Tests for variation over groups in survival data. *Journal of the American Statistical Association*, 90(429), 198–203.

Henderson, R. & Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(2), 367–379.

Holt, J. D. & Prentice, R. L. (1974). Survival analyses in twin studies and matched pair experiments. *Biometrika*, 61(1), 17–30.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2), 387–396.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer: New York.

Huber, P., Ronchetti, E., & Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(4), 893–908.

Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 48(3), 795–806.

Lango, M. N., Egleston, B., Ende, K., Feigenberg, S., D'Ambrosio, D. J., Cohen, R. B., Ahmad, S., Nicolaou, N., & Ridge, J. A. (2010). Impact of neck dissection on long-term feeding tube dependence in patients with head and neck cancer treated with primary radiation or chemoradiation. *Head & Neck*, 32(3), 341–347.

Lee, E. W., Wei, L. J., & Amato, D. A. (1992). Cox-type regression analysis for large number of small groups of correlated failure time observations. In J. P. Klein & P. K. Goel (Eds.), *Survival Analysis: State of the Art*. Kluwer: Dordrecht.

Li, Y. & Lin, X. (2006). Semiparametric normal transformation models for spatially correlated survival data. *Journal of the American Statistical Association*, 101(474), 591–603.

Liang, K.-Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1), 3–40.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine*, 13(21), 2233–2247.

Lin, D. Y. & Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84(408), 1074–1078.

Lin, D. Y. & Wei, L. J. (1991). Goodness-of-fit tests for the general cox regression model. *Statistica Sinica*, 1, 1–17.

Lorino, T., Sanaa, M., Robin, S., & Daudin, J. (2004). Comparison of semiparametric regression models for correlated survival data using simulations. *Communications in Statistics*, 33(8), 1975–1991.

Machtay, M., Moughan, J., Trotti, A., Garden, A. S., Weber, R. S., Cooper, J. S., Forastiere, A., & Ang, K. K. (July 20, 2008). Factors associated with severe late toxicity after concurrent chemoradiation for locally advanced head and neck cancer: An rtog analysis. *Journal of Clinical Oncology*, 26(21), 3582–3589.

McGilchrist, C. A. (1993). Reml estimation for survival models with frailty. *Biometrics*, 49(1), 221–225.

McGilchrist, C. A. & Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, 47(2), 461–466.

Moustaki, I. & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65(3), 391–411.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*, 22(2), 712–731.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics*, 23(1), 182–198.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132.

Nguyen, N. P., Frank, C., C, M. C., Vos, P., Smith, H. J., Nguyen, P. D., Martinez, T., Karlsson, U., Dutta, S., Lemanski, C., Nguyen, L. M., & Sallah, S. (2009). Analysis of factors influencing aspiration risk following chemoradiation for oropharyngeal cancer. *British Journal of Radiology*, 82(980), 675–680.

Nielsen, G. G., Gill, R. D., Andersen, P. K., & Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19(1), 25–43.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406), 487–493.

Othus, M. (2009). PhD thesis, Harvard University.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics*, 26(1), 183–214.

Pignon, J., Bourhis, J., Domenge, C., & Design, L. (2000). Chemotherapy added to locoregional treatment for head and neck squamous-cell carcinoma: three meta-analyses of updated individual data. *The Lancet*, 355(9208), 949 – 955.

Pocock, S. J., Geller, N. L., & Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, 43(3), 487–498.

Posner, M. R., Hershock, D. M., Blajman, C. R., Mickiewicz, E., Winquist, E., Gorbounova, V., Tjulandin, S., Shin, D. M., Cullen, K., Ervin, T. J., Murphy, B. A., Raez, L. E., Cohen, R. B., Spaulding, M., Tishler, R. B., Roth, B., Viroglio, R. d. C., Venkatesan, V., Romanov, I., Agarwala, S., Harter, K. W., Dugan, M., Cmelak, A., Markoe, A. M., Read, P. W., Steinbrenner, L., Colevas, A. D., Norris, C. M., & Haddad, R. I. (2007). Cisplatin and fluorouracil alone or with docetaxel in head and neck cancer. *New England Journal of Medicine*, 357(17), 1705–1715.

Prentice, R. L. & Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, 79(3), 495–512.

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (1992). *Numerical Recipes in C: The Art of Scientic Computing*. Cambridge University Press, second edition.

Regan, M. M. & Catalano, P. J. (1999). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics*, 55(3), 760–768.

Roy, J. & Lin, X. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, 56(4), 1047–1054.

Sammel, M. D. & Ryan, L. M. (1996). Latent variable models with fixed effects. *Biometrics*, 52(2), 650–663.

Sammel, M. D., Ryan, L. M., & Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3), 667–678.

Seibel, N. L., Steinherz, P. G., Sather, H. N., Nachman, J. B., DeLaat, C., Ettinger, L. J., Freyer, D. R., Mattano, Leonard A., J., Hastings, C. A., Rubin, C. M., Bertolone, K., Franklin, J. L., Heerema, N. A., Mitchell, T. L., Pyesmany, A. F., La, M. K., Edens, C., & Gaynon, P. S. (2008). Early postinduction intensification therapy improves survival for children and adolescents with high-risk acute lymphoblastic leukemia: a report from the Children's Oncology Group. *Blood*, 111(5), 2548–2555.

Senn, S. (1998). Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine*, 17, 1753–1765.

Shih, J. H. & Louis, T. A. (1995). Assessing gamma frailty models for clustered failure time data. *Lifetime Data Analysis*, 1, 205–220.

Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), pp. 201–292.

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer: New York.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3), 439–454.

Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408), 1065–1073.